Advance in Statistical Theory and Methods for Social Sciences

Ying Lu

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistics and Operations Research.

Chapel Hill
June 4, 2009

Approved by

Edward Carlstein

Jianqing Fan

Chuanshu Ji

Steve Marron

Zhengyuan Zhu

ABSTRACT

Ying Lu: Advance in Statistical Theory and Methods for Social Sciences
(Under the direction of Jianqing Fan)

This dissertation includes three papers. In the first paper, a new statistical procedure is proposed to analyze verbal autopsy data. Verbal autopsy procedures are widely used for estimating cause-specific mortality in areas without medical death certifications. We show that the problem of estimating cause-specific mortality rate can be directly solved using the distribution of symptoms that is available from the population verbal autopsy survey and the cause-specific distribution of symptoms that can be obtained from hospital data. To solve this deconvolution problem, we offer an optimization procedure that is stable and easy to compute. Through empirical analyses in data from China and Tanzania, we illustrate the accuracy of this approach.

In the second paper, we focus on the analysis of roll call and vote records data to legislative and judicial voting behaviors. Ideal point estimation is an important tool to analyze this type of data. We introduce a hierarchical ideal point estimation framework that directly models complex voting behaviors based on the characteristics of the political actors and the votes they cast. Bayesian MCMC algorithms are proposed to estimate the proposed hierarchical models. Through simulations and empirical examples we show that this framework holds good promise for resolving many unsettled issues, such as the multi-dimensional aspects of ideology, and the effects of political parties.

In the third paper, we address the issue of variable selection in linear mixed effect models. Mixed effect models are fundamental tools for the analysis of longitudinal data, panel data and cross-sectional data. However, the complex nature of these models has made variable selection and parameter estimation a challenging problem. In this paper, we propose a simple iterative procedure that estimates and selects fixed and random effects for linear mixed models. In particular, we propose to utilize the partial consistency property of the random effect coefficients and select groups of random effects simultaneously via a data-oriented penalty function. We show that the proposed method is a consistent variable selection procedure and possesses the Oracle properties. Simulation studies and a real data analysis are also conducted to empirically examine the performance of this procedure.

ACKNOWLEDGEMENT

First, I owe Dr. Jianqing Fan much for his support and continuing inspiration. Without his encouragement, I could not possibly finish this dissertation. His keen view of statistics has fundamentally shaped my career choices and broadened and deepened my perspectives as an applied statistician. I feel fortunate to study under his advice as a graduate student.

I would like to express my deep gratitude to the members of the dissertation committee, Dr. Chuanshu Ji, Dr. Ed Carlstein, Dr. Steve Marron and Dr. Zhengyuan Zhu. Your patience with me and generosity in accommodating my schedules has made it possible for me to finish this dissertation remotely in Colorado. I also thank Dr. Andrew Nobel for helping me return to return to UNC-Chapel Hill after a long period of leave.

Much of my gratitude also goes to my collaborators, Gary King, Xiaohui Wang and Heng Peng. The many discussions and email correspondences with them formed the foundation of this dissertation.

I would also like to thank the staff members at the Department of Statistics and Operations Research, Ms. Charlotte Rogers, Alison Kieber and Elisabeth Moffitt-Johnson.

Last but not least, much love to my husband. Brent, thanks for being here for me all the time, no matter what.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This dissertation is a compilation of three research projects each of which addresses an important issue in social science methodology.

Verbal autopsy procedures are widely used for estimating cause-specific mortality in areas without medical death certification. Data on symptoms reported by caregivers along with the cause of death are collected from a medical facility, and the cause-of-death distribution is estimated in the population where only symptom data are available. Current approaches analyze only one cause at a time, involve assumptions judged difficult or impossible to satisfy, and require expensive, time consuming, or unreliable physician reviews, expert algorithms, or parametric statistical models. By generalizing current approaches to analyze multiple causes, we show how most of the difficult assumptions underlying existing methods can be dropped. These generalizations also make physician review, expert algorithms, and parametric statistical assumptions unnecessary. With theoretical results, and empirical analyses in data from China and Tanzania, we illustrate the accuracy of this approach. While no method of analyzing verbal autopsy data, including the more computationally intensive approach offered here, can give accurate estimates in all circumstances, the procedure offered is conceptually simpler, less expensive, more general, as or more replicable, and easier to use in practice than existing approaches. We also show how our focus on estimating aggregate proportions, which are the quantities of primary interest in verbal autopsy studies, may also greatly reduce the assumptions necessary, and thus improve the performance of, many individual classifiers in this and other areas. As a companion to this paper, we also offer easy-to-use software that implements the methods discussed herein.

Ideal point estimation is an important tool for political scientists to study legislator's voting behaviors and to assess their political preferences. In traditional statistical models of ideal point estimation, individuals are unrealistically assumed to make decisions independently from each other, and to make each decision inde-

pendently from other decisions. When such assumptions do not hold, the parameters estimated from the traditional ideal point estimation model tend to be biased and inefficient. Moreover, failing to address these issues has limited the ideal point research from understanding important topics such as party influence, period effect and the multidimensional nature of political ideology. In this paper, we propose a hierarchical ideal point estimation framework that directly models inter-individual and intra-individual correlations in legislative behaviors via random effects and/or fixed effects. Under this framework, modelers can define clusters of individuals (*allysets*), clusters of bills (*votesets*), and/or voting blocks (*tactsets*) based on the characteristics of the legislators and the votes they cast. The effects and the significance of these ex ante clusters can then be assessed statistically. Such setup entails a substantively intuitive and methodologically coherent approach to test political theory of legislative behaviors. Through simulations and empirical examples of the legislative behaviors of the US supreme court and the House of Representatives, we show that the proposed framework holds good promise for resolving many unsettled issues in ideal point estimation. In addition, the proposed framework can be readily extended to a more general family of 2-parameter Rasch model with applications in other fields. As a companion to this paper, we offer an easy-to-use R package with C code that implements the methods discussed herein.

Lastly, in this dissertation, the problem of variable selection for linear mixed effects model will be studied. Mixed effect models are fundamental tools for the analysis of longitudinal data, panel data and cross-sectional data. They are widely used by various fields of social sciences, medical and biological sciences. However, the complex nature of these models has made variable selection and parameter estimation a challenging problem. In this paper, we propose a simple iterative procedure that estimates and selects fixed and random effects for linear mixed models. In particular, we propose to utilize the partial consistency property of the random effect coeffi-

cients and select groups of random effects simultaneously via a data-oriented penalty function (the smoothly clipped absolute deviation penalty function). We show that the proposed method is a consistent variable selection procedure and possesses some oracle properties. Simulation studies and a real data analysis are also conducted to empirically examine the performance of this procedure.

Chapter 2

Verbal Autopsy Methods with Multiple Causes of Death

2.1 Introduction

National and international policymakers, public health officials, and medical personnel need information about the global distribution of deaths by cause in order to set research goals, budgetary priorities, and ameliorative policies. Yet, only 23 of the world's 192 countries have high quality death registration data, and 75 have no cause-specific mortality data at all (Mathers et al., 2005). Even if we include data of dubious quality, less than a third of the deaths that occur worldwide each year have a cause certified by medical personnel (Lopez et al., 2000).

Verbal autopsy is a technique "growing in importance" (Sibai et al., 2001) for estimating the cause-of-death distribution in populations without vital registration or other medical death certification. It involves collecting information about symptoms (including signs and other indicators) from the caretakers of each of a randomly selected set of deceased in some population of interest, and inferring the cause of

death. Inferences in these data are extrapolated either by physicians from their prior experiences or by statistical analysis of a second data set from a nearby hospital where information on symptoms from caretakers as well as validated causes of death are available.

Verbal autopsy studies are now widely used throughout the developing world to estimate cause-specific mortality, and are increasingly being used for disease surveillance and sample registration (Setel et al.,2005). Verbal autopsy is used on an ongoing basis and on a large scale in India and China, and in 36 demographic surveillance sites around the world (Soleman, Chandramohan and Shibuya, 2005). The technique has also proven useful in studying risk factors for specific diseases, infectious disease outbreaks, and the effects of public health interventions (Anker, 2003; Pacque- Margolis et al., 1990; Soleman, Chandramohan and Shibuya, 2006).

Until now, the most commonly used method has been physician review of symptoms with no additional validation sample. This approach can be expensive as it involves approximately three physicians, each taking 20-30 minutes to review symptoms and classify each death. To reduce the total time necessary, more physicians can be hired and work in parallel. Because judgments by these doctors are highly sensitive to their priors (when a Kansas doctor hears "fever and vomiting," malaria would not be her first thought), physicians need to come from local areas. This can pose difficult logistical problems because physicians in these areas are typically in very short supply, as well as serious ethical dilemmas since doctors are needed in the field for treating patients. Physician review also poses scientific problems since, although scholars have worked hard at increasing inter-physician reliability for individual studies, the cross-study reliability of this technique has remained low. Attempts to formalize physician reviews via expert-created deterministic algorithms are reliable by design, but appear to have lower levels of validity, in part because many diseases are not modeled explicitly and too many decisions need to be made.

Inferences from verbal autopsy data would thus seem ripe for adding to the growing list of areas where radically empirical approaches imbued with the power of modern statistics dominate human judgments by local experts (Dawes, Faust and Meehl, 1989). Unfortunately, the parametric statistical modeling that has been used in this area (known in the field as "data-derived techniques") have suffered from low levels of agreement with verified causes of death and are complicated for large numbers of causes. In practice, the choice of model has varied with almost every application. We attempt to rectify this situation.

In this article, we describe the current verbal autopsy approaches and the not always fully appreciated assumptions underlying them. We show that a key problem researchers have in satisfying most of the assumptions in real applications can be traced to the constraint existing methods impose by requiring the analysis of only one cause of death at a time. We generalize current methods to allow many causes of death to be analyzed simultaneously. This simple generalization turns out to have some considerable advantages for practice, such as making it unnecessary to conduct expensive physician reviews, specify parametric statistical models that predict the cause of death, or build elaborate expert algorithms. Although the missing (cause of death) information guarantees that verbal autopsy estimates always have an important element of uncertainty, the new approach offered here greatly reduces the unverified assumptions necessary to draw valid inferences. As a companion to this article, we are making available easy-to-use, free, and open source software that implements all our procedures.

The structure of the inferential problem we study can also be found in application areas fairly distant from our verbal autopsy applications. Some version of the methods we discuss may be of use in these areas as well. For example, a goal of paleodemography is to estimate the age distribution in a large sample of skeletons from measurements of their physical features by using a small independent reference group

where validated ages are available and skeletal features are also measured (Hoppa and Vaupel, 2002). Our methods seem to have already proven useful for estimating the proportion of text documents in each of a set of given categories, using a smaller reference set of text documents hand coded into the same categories (Hopkins and King, 2007). Also, as we show in Section 2.8, the methods introduced here imply that individual level classifiers can greatly reduce the assumptions necessary for accurate generalization to test sets with different distributional characteristics.

2.2 Data Definitions and Inferential Goals

Denote the cause of death j (for possible causes $j = 1, \ldots, J$) of individual i as $D_i = j$. Bereaved relatives or caretakers are asked about each of a set of symptoms (possibly including signs or other indicators) experienced by the deceased before death. Each symptom k (for possible symptoms $k = 1, \ldots, K$) is reported by bereaved relatives to have been present, which we denote for individual i as $S_{ik} = 1$, or absent, $S_{ik} = 0$. We summarize the set of symptoms reported about an individual death, $\{S_{i1}, \ldots, S_{iK}\}$, as the vector $\boldsymbol{S}_i$. Thus, the cause of death D_i is one variable with many possible values, whereas the symptoms $\boldsymbol{S}_i$ constitute a set of variables, each with a dichotomous outcome.

Data come from two sources. The first is a hospital or other validation site, where both $\boldsymbol{S}_i$ and D_i are available for each individual i ($i = 1, \ldots, n$). The second is the community or some population about which we wish to make an inference, where we observe $\boldsymbol{S}_\ell$ (but not D_ℓ) for each individual ℓ ($\ell = 1, \ldots, L$). Ideally, the second source of data constitutes a random sample from a large population of interest, but it could also represent any other relevant target group.

The quantity of interest for our analysis is $P(D)$, the distribution of cause-specific mortality in the population. Public health scholars are not normally interested in

the cause of death D_ℓ of any particular individual in the population (although some current methods require estimates of these as intermediate values to compute $P(D)$). They are sometimes also interested in the cause of death distribution for subgroups, such as age, sex, region, or condition. We return to the implications of our approach for individual level classifiers in Section 2.8.

The difficulty of verbal autopsy analyses is that the population cause of death distribution is not necessarily the same in the hospital where D is observed. In addition, researchers often do not sample from the hospital randomly, and instead over-sample deaths due to causes that may be rare in the hospital. Thus, in general, the cause of death distribution in our two samples cannot be assumed to be the same: $P(D) \neq P^h(D)$.

Since symptoms are *consequences* of the cause of death, the data generation process has a clear ordering: Each disease or injury $D = j$ produces some symptom profiles (sometimes called "syndromes" or values of $\boldsymbol{S}$) with higher probability than others. We represent these conditional probability distributions as $P^h(\boldsymbol{S}|D)$ for data generated in the hospital and $P(\boldsymbol{S}|D)$ in the population. Thus, since the distribution of symptom profiles equals the distribution of symptoms given deaths weighted by the distribution of deaths, the symptom distribution will not normally be observed to be the same in the two samples: $P(\boldsymbol{S}) \neq P^h(\boldsymbol{S})$.

Whereas $P(D)$ is a multinomial distribution with J outcomes, $P(\boldsymbol{S})$ may be thought of as either a multivariate distribution of K binary variables or equivalently as a univariate multinomial distribution with 2^K possible outcomes, each of which is a possible symptom profile. We will usually use the 2^K representation.

2.3 Current Estimation Approaches

The most widely used current method for estimating cause of death distributions in verbal autopsy data is physician review. What appears to be the best practice among the current statistical approaches used in the literature is the following multi-stage estimation strategy.

1. Choose a cause of death, which we here refer to as cause of death $D = 1$, apply the remaining steps to estimate $P(D = 1)$, and then repeat for each additional cause of interest (changing 1 to 2, then 3, etc).

2. Using hospital data, develop a method of using a set of symptoms $\boldsymbol{S}$ to create a prediction for D, which we label $\hat{D}$ (and which takes on the value 1 or not 1). Some do this directly using informal, qualitative, or deterministic prediction procedures, such as physician review or expert algorithms. Others use formal statistical prediction methods (called "data-derived algorithms" in the verbal autopsy literature), such as logistic regression or neural networks, which involve fitting $P^h(D|\boldsymbol{S})$ to the data and then turning it into a 0/1 prediction for an individual. Typically this means that if the estimate of $P^h(D = 1|\boldsymbol{S})$ is greater than 0.5, set the prediction as $\hat{D} = 1$ and otherwise set $\hat{D} \neq 1$. Of course, physicians and those who create expert algorithms implicitly calculate $P^h(D = 1|\boldsymbol{S})$, even if they never do so formally.

3. Using data on the set of symptoms for each individual in the community, $\boldsymbol{S}_\ell$, and the same prediction method fit to hospital data, $P^h(D_\ell = 1|\boldsymbol{S}_\ell)$, create a prediction $\hat{D}_\ell$ for all individuals sampled in the community ($\ell = 1, \ldots, L$) and average them to produce a preliminary or "crude" estimate of the prevalence of the disease of interest, $P(\hat{D} = 1) = \sum_{\ell=1}^{L} \hat{D}_\ell / L$.

4. Finally, estimate the *sensitivity*, $P^h(\hat{D} = 1|D = 1)$, and *specificity*, $P^h(\hat{D} \neq$

$1|D \neq 1)$, of the prediction method in hospital data and use it to "correct" the crude estimate and produce the final estimate:

$$P(D=1) = \frac{P(\hat{D}=1) - [1 - P^h(\hat{D} \neq 1|D \neq 1)]}{P^h(\hat{D}=1|D=1) - [1 - P^h(\hat{D} \neq 1|D \neq 1)]} \tag{2.1}$$

This correction, sometimes known as "back calculation", was first described in the verbal autopsy literature by Kalter (1992, Table 1) and originally developed for other purposes by Levy and Kass (1970). The correction is useful because the crude prediction, $P(\hat{D}=1)$, can be inaccurate if sensitivity and specificity are not 100%.

A variety of creative modifications of this procedure have also been tried (Chandramohan, Maude, Rodrigues and Hayes, 1994). These include meta-analyses of collections of studies (Morris, Black and Tomaskovic, 2003), different methods of estimating $\hat{D}$, many applications with different sets of symptoms and different survey instruments (Soleman, Chandramohan and Shibuya, 2006), and other ways of combining the separate analyses from different diseases (Quigley et al., 2000; Boulle, Chandramohan and Weller, 2001).

2.4 Assumptions Underlying Current Practice

The method described in Section 3.2 makes three key assumptions that we now describe. Then in the following section, we develop a generalized approach that reduces our reliance on the first assumption and renders the remaining two unnecessary.

The first assumption is that the sensitivity and specificity of $\hat{D}$ estimated from the hospital data are the same as that in the population:

$$\begin{aligned} P(\hat{D}=1|D=1) &= P^h(\hat{D}=1|D=1) \\ P(\hat{D} \neq 1|D \neq 1) &= P^h(\hat{D} \neq 1|D \neq 1). \end{aligned} \tag{2.2}$$

The literature contains much discussion of this assumption, the variability of estimates of sensitivity and specificity across sites, and good advice about controlling their variability (Kalter, 1992).

A less well known but worrisome aspect of this first assumption arises from the choice of analyzing the J-category death variable as if it were a dichotomy. Because of the composite nature of the aggregated $D \neq 1$ category of death, we must assume that what makes up this composite is the same in the hospital and population. If it is not, then the required assumption about specificity (i.e., about the accuracy of estimation of this composite category) cannot hold in the hospital and population, even if sensitivity is the same. In fact, satisfying this assumption is more difficult than may be generally understood. To make this point, we begin with the decomposition of specificity, offered by Chandramohan, Setel and Quigley (2001) (see also Maude and Ross, 1997), as one minus the sum of the probability of different misclassifications times their respective prevalences:

$$P(\hat{D} \neq 1|D \neq 1) = 1 - \sum_{j=2}^{J} P(\hat{D} = 1|D = j)\frac{P(D = j)}{P(D \neq 1)}, \tag{2.3}$$

which emphasizes the composite nature of the $D \neq 1$ category. Then we ask: *under what conditions can specificity in the hospital equal that in the population if the distribution of cause of death differs?* The mathematical condition can be easily derived by substituting (2.3) into each side of the second equation of (2.2) (and simplifying by dropping the "1−" on both sides):

$$\sum_{j=2}^{J} P(\hat{D} = 1|D = j)\frac{P(D = j)}{P(D \neq j)} = \sum_{j=2}^{J} P^h(\hat{D} = 1|D = j)\frac{P^h(D = j)}{P^h(D \neq j)} \tag{2.4}$$

If this equation holds, then this first assumption holds. And if $J = 2$, this equation reduces to the first line of (2.2) and so, in that situation, the assumption is unproblematic.

However, for more than two diseases specificity involves a composite cause of death category. We know that the distribution of causes of death (the last factor on each side of Equation 2.4) differs in the hospital and population by design, and so the equation can hold only if a miraculous mathematical coincidence holds, whereby the probability of misclassifying each cause of death as the first cause occurs in a pattern that happens to cancel out differences in the prevalence of causes between the two samples. For example, this would not occur according to any theory or observation of mortality patterns offered in the literature. Verbal autopsy scholars recognize that some values of sensitivity and specificity are impossible when (2.1) produces estimates of $P(D = 1)$ greater than one. They then use information to question the values of, or modify, estimates of sensitivity and specificity, but the problem is not necessarily due to incorrect estimates of these quantities and could merely be due to the fact that the procedure requires assumptions that are impossible to meet. In fact, *as the number of causes of death increase, the required assumption can only hold if sensitivity and specificity are each 100%*, which we know does not describe real data.[1]

The second assumption is that the (explicit or implicit) model underlying the prediction method used in the hospital must also hold in the population: $P(D|\boldsymbol{S}) = P^h(D|\boldsymbol{S})$. For example, if logistic regression is the prediction method, we make this assumption by taking the coefficients estimated in hospital data and using them to multiply by symptoms collected in the population to predict the the cause of death in the population. This is an important assumption, but not a natural one since the data generation process is the reverse: $P(\boldsymbol{S}|D)$. And most importantly, even if the identical

[1]The text describes how this first assumption can be met by discussing specificity only with respect to cause of death 1. In the general case, (2.4) for all causes requires satisfying $\sum_j P(\hat{D} \neq j|D \neq j) - (J-2) = \sum_j [P(\hat{D} \neq j|D \neq j) + P(\hat{D} = j|D = j)]P(D = j)$. For small $J > 2$, this will hold only if a highly unlikely mathematical coincidence occurs; for large J, this condition is not met in general unless sensitivity and specificity is 1 for all j.

data generation process held in the population and hospital, $P(\boldsymbol{S}|D) = P^h(\boldsymbol{S}|D)$, we would still have no reason to believe that $P(D|\boldsymbol{S}) = P^h(D|\boldsymbol{S})$ holds. The assumption might hold by luck, but coming up with a good reason why we should believe it holds in any real case seems unlikely.

This problem is easy to see by generating data from a regression model with D as the explanatory variable and $\boldsymbol{S}$ as the simple dependent variable, and then regressing $\boldsymbol{S}$ on D: Unless the regression fits perfectly, the coefficients from the first regression do not determine those in the second. Similarly, when Spring comes, we are much more likely to see many green leaves; but visiting the vegetable section of the supermarket in the middle of the winter seems unlikely to cause the earth's axis to tilt toward the sun. Of course, it just *may* be that we can find a prediction method for which $P(D|\boldsymbol{S}) = P^h(D|\boldsymbol{S})$ holds, but knowing whether it does or even having a theory about it seems unlikely. It is also *possible*, with a small number of causes of death, that the sensitivity and specificity for the wrong model fit to hospital data could by chance be correct when applied to the population, but it is hard to conceive of a situation when we would know this ex ante. This is especially true given the issues with the first assumption: the fact that the composite $D \neq 1$ category is by definition different in the population and hospital implies that different symptoms will be required predictors for the two models, hence invalidating this assumption.

A final problem with the current approach is that the multi-stage procedure estimates $P(D = j)$ for each j separately, but for the ultimate results to make any sense the probability of a death occurring due to some cause must be 100%: $\sum_{j=1}^{J} P(D = j) = 1$. This can happen if the standard estimation method is used, but it will hold only by chance.

2.5 An Alternative Approach

The key problem underlying the veracity of each of the assumptions in Section 2.4 can be traced to the practice of sequentially dichotomizing the J-category cause of death variable. In analyzing the first assumption, we learn that specificity cannot be equal in hospital and population data as the number of causes that make up the composite residual category gets large. In the second assumption, the practice of collapsing the relationship between $\boldsymbol{S}$ and D into a dichotomous prediction, $\hat{D}$, requires making assumptions opposite to the data generation process and either a sophisticated statistical model, or an expensive physician review or set of expert algorithms, to summarize $P(D|S)$. And finally, the estimated cause of death probabilities do not necessarily sum to one in the existing approach precisely because D is dichotomized in multiple ways and each dichotomy is analyzed separately.

Dichotomization has been used in each case to simplify the problem. However, we show in this section that most aspects of the assumptions with the existing approach are unnecessary once we treat the J-category cause of death variable as having J categories. Moreover, it is simpler conceptually than the current approach. We begin by *reformulating* the current approach so it is more amenable to further analysis and then *generalizing* it to the J-category case.

Reformulation Under the current method's assumption that sensitivity and specificity are the same in the hospital and population, we can rearrange the back-calculation formula in (2.1) as

$$P(\hat{D}=1) = P(\hat{D}=1|D=1)P(D=1) + P(\hat{D}=1|D\neq 1)P(D\neq 1). \tag{2.5}$$

and rewrite (2.5) in equivalent matrix terms as

$$\underset{2\times 1}{P(\hat{D})} = \underset{2\times 2}{P(\hat{D}|D)}\underset{2\times 1}{P(D)} \tag{2.6}$$

where the extra notation indicates the dimension of the matrix or vector. So $P(\hat{D})$ and $P(D)$ are now both 2×1 vectors, and have elements $[P(\hat{D} = 1), P(\hat{D} \neq 1)]'$ and $[P(D = 1), P(D \neq 1)]'$, respectively; and $P(\hat{D}|D)$ is a 2×2 matrix where

$$\underset{2\times 2}{P(\hat{D}|D)} = \begin{pmatrix} P(\hat{D} = 1|D = 1) & P(\hat{D} = 1|D \neq 1) \\ P(\hat{D} \neq 1|D = 1) & P(\hat{D} \neq 1|D \neq 1) \end{pmatrix}.$$

Whereas (2.1) is solved for $P(D = 1)$ by plugging in values for each term on the right side, (2.6) is solved for $P(D)$ by linear algebra. Fortunately, the linear algebra required is simple and well known from the least squares solution in linear regression. We thus recognize $P(\hat{D})$ as taking the role of a "dependent variable," $P(\hat{D}|D)$ as two "explanatory variables," and $P(D)$ as the coefficient vector to be solved for. Applying least squares yields an estimate of $P(D)$, the first element of which, $P(D = 1)$, is exactly the same as that in Equation 2.1. Thus far, only the mathematical representation has changed; the assumptions, intuitions, and estimator remain identical to the existing method described in Section 3.2.

Generalization The advantage of switching to matrix representations is that they can be readily generalized, which we do now in two important ways. First, we drop the modeling necessary to produce the cause of death for each individual $\hat{D}$, and use $\boldsymbol{S}$ in its place directly. And second, we do not dichotomize D and instead treat it as a full J-category variable. We implement both generalizations via a matrix expression that is the direct analogue of (2.6):

$$\underset{2^K\times 1}{P(\boldsymbol{S})} = \underset{2^K\times J}{P(\boldsymbol{S}|D)} \underset{J\times 1}{P(D)} \tag{2.7}$$

The quantity of interest in this expression remains $P(D)$. Although we use the better nonparametric estimation methods (described in the appendix), we could in principle estimate $P(\boldsymbol{S})$ by direct tabulation, by simply counting the fraction of people in the

population who have each symptom profile. Since we do not observe and cannot directly estimate $P(\boldsymbol{S}|D)$ in the community (because D is unobserved), we estimate it from the hospital and assume $P(\boldsymbol{S}|D) = P^h(\boldsymbol{S}|D)$. We estimate $P^h(\boldsymbol{S}|D = j)$ for each cause of death j the same way as we do for $P(\boldsymbol{S})$.

The only assumption required for connecting the two samples is $P(\boldsymbol{S}|D) = P^h(\boldsymbol{S}|D)$, which is natural as it directly corresponds to the data generation process. We do not assume that $P(\boldsymbol{S})$ and $P^h(\boldsymbol{S})$ are equal, $P(D)$ and $P^h(D)$ are equal, or $P(D|\boldsymbol{S})$ and $P^h(D|\boldsymbol{S})$ are equal. In fact, prediction methods for estimating $P(D|\boldsymbol{S})$ or $\hat{D}$ are entirely unnecessary here, and so unlike the current approach, we do not require parametric statistical modeling, physician review, or expert algorithms.

We solve Equation 2.7 for $P(D)$ directly. This can be done conceptually using least squares. That is, $P(\boldsymbol{S})$ takes the role of a "dependent variable," $P(\boldsymbol{S}|D)$ takes the role of a matrix of J "explanatory variables," each column corresponding to a different cause of death, and $P(D)$ is the "coefficient vector" with J elements for which we wish to solve. We also modify this procedure to ensure that the estimates of $P(D)$ are each between zero and one and together sum to one by changing least squares to constrained least squares (see the Appendix).

Although producing estimates from this expression involves some computational complexities, this is a single equation procedure that is conceptually far simpler than current practice. As described in Section 3.2, the existing approach requires four steps, applied sequentially to each cause of death. In contrast, estimates from our proposed alternative only require understanding each term in Equation 2.6 and solving for $P(D)$.

2.6 Illustrations in Data from China and Tanzania

Since deaths are not observed in populations for which verbal autopsy methods are used, realistic validation of any method is, by definition, difficult or impossible (Gajalakshmi and Peto, 2004). We attempt to validate our method in two separate ways in data from China and Tanzania.

China We begin with an analysis of 2,822 registered deaths from hospitals in urban China collected and analyzed by Alan Lopez and colleagues (see, most recently, Yang et al., 2005). Thirteen causes of death were coded, and 56 (yes or no) symptoms were elicited from caretakers. We conducted three separate analyses with these data. We designed the first test to meet the assumptions of our method by randomly spliting these data into halves. Although all these data were collected in hospitals, where we observe both $\boldsymbol{S}$ and D, we label the first random set "hospital data," for which we use both $\boldsymbol{S}$ and D, and the second "population data," for which we *only* use $\boldsymbol{S}$ during estimation. We emulate an actual verbal autopsy analysis by using these data to estimate the death frequency distribution, $P(D)$, in the "population data." Finally, for validation, we unveil the actual cause of death variable for the "population data" that were set aside during the analysis and compare it to our estimates.

The estimates appear in the top panel of the left graph of Figure 2.1, which plots on the horizontal axis a direct sample estimate — the proportion of the sample from the population dying from each of 13 causes — and on the vertical axis an estimate from our verbal autopsy method. (This direct estimator is not normally feasible in verbal autopsy studies because of the impossibility of obtaining medically verified cause of death data in the community.) Since both are sample-based estimates, and thus both are measured with error, if our method predicted perfectly, all points would fall approximately on the 45 degree line. Clearly, the fit of our estimates to

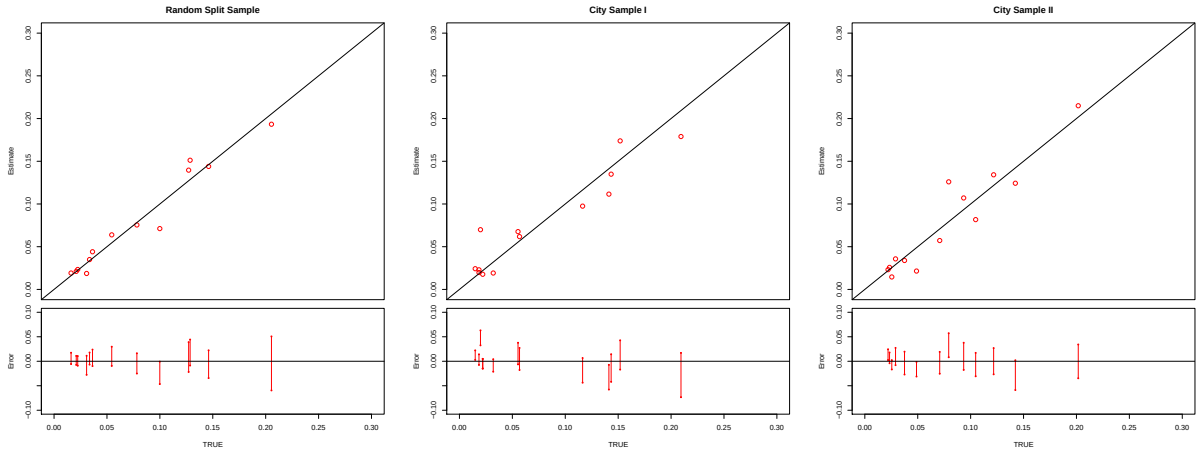

Figure 2.1: Validation in China. A direct estimate of cause-specific mortality is plotted horizontally in the top panel by the estimate from our method plotted vertically for randomly split data (top left) and for predictions of one set of hospitals to another (the two top right graphs). The bottom panel of each graph contains 95% confidence intervals of the difference between our estimator and the direct estimate, both of which are measured with error; almost all of these vertical lines cross the zero difference point marked by a horizontal line.

the direct estimates of the truth is fairly close, with no clear pattern in deviations from the line. The bottom panel of this graph portrays the difference between our estimates and the direct sample estimates, along with a 95% confidence interval for the difference. Almost all confidence intervals of the errors cover no difference (portrayed as a horizontal line), which indicates approximately accurate coverage.

For a more stringent test of our approach, we split the same sample into 1,409 observations from hospitals in three cities (Beijing, Chengdu, and Wuhan) and 1,413 observations from hospitals in another three cities (Haierbin, Guangzhou, and Shanghai). We then let each group takes a turn playing the role of the "population" sample (with known cause of death that we use only for validation) and the other as the

actual hospital sample. These are more difficult tests of our method than would be necessary in practice, since researchers would normally collect hospital data from a facility physically closer to, part of, and more similar to the population to which they wish to infer.

The right two graphs in Figure 2.1 give results from this test in the same format as for the random split on the left. The middle graph estimates the cause of death distribution of our first group of sample cities from the second group, whereas the right graph does the reverse. The fit between the directly estimated true death proportions and our estimates in both is slightly worse than for the left graph, where our assumptions were true by construction, but predictions in both are still excellent. Again, almost all of the 95% confidence intervals for the difference between our estimator and the direct sample estimate cross the zero line (see the bottom of each graph).

Tanzania We also analyze cause-specific adult mortality from a verbal autopsy study in Tanzania (see Setel et al., 2006). The data include 1,261 hospital deaths and 282 deaths from the general population, about which 51 symptoms questions and 13 causes of death were collected. The unusual feature of these data is that all the population deaths have medically certified causes, and so we can set aside that information and use it to validate our approach. We again use $\boldsymbol{S}$ and D from the hospital and $\boldsymbol{S}$ from the population and attempt to estimate $P(D)$ in the population, using D from the population only for validation after the estimation is complete.

The results appear in Figure 2.2 in the same format as the China data. We constructed randomly split data on the left and an actual prediction to the community for the graph on the right. The results are similar to that in China, where the point estimates appear roughly spread around the 45° line, indicating, in this very different context, that the fit is approximately as good — and again better for the random

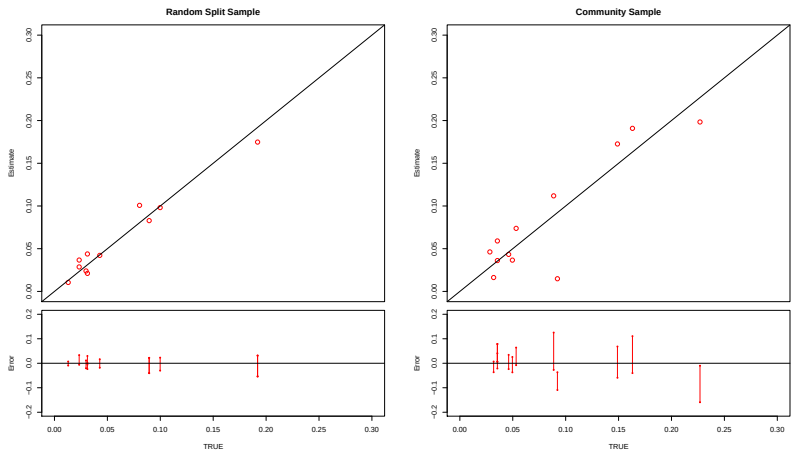

Figure 2.2: Validation in Tanzania. Each graph plots the (normally unknown) direct estimate of cause-specific mortality horizontally and estimates from our method vertically. This is done for data based on a random split, where our assumptions are true by construction, on the left and for predictions of the community sample based on hospital sample on the right.

split than the actual forecast. The confidence intervals of the differences between the direct estimate and our estimate, in the bottom panel, are larger than for the China data due to the smaller target population used to estimate $P(\boldsymbol{S})$, but almost all the intervals cross zero.

The variance of the direct sampling estimator, $\bar{D}_j$, is approximately $\bar{D}_j(1-\bar{D}_j)/n$, and thus varies with category size. Uncertainty estimates from our approach are computed by bootstrapping, and of course also vary by category size. The 95% confidence interval from our estimator is on average across categories 50% wider than the direct sampling estimator in the China data and 25% wider in the Tanzania data. Obviously, the reason verbal autopsy procedures are necessary is that direct sampling estimates of the cause of death in the population are unobtainable, and so these numbers summarize the necessary costs incurred for this lack of information.

Of course, compared to the huge costs of complete national vital registration systems, this is a trival difference.

2.7 Interpretation

We offer five interpretations of our approach. First, since $\boldsymbol{S}$ contains K dichotomous variables and thus 2^K symptom profiles, $P(\boldsymbol{S})$ and $P(\boldsymbol{S}|D)$ have 2^K rows, which take the role of "observations" in this linear expression. By analogy to linear regression, where more observations make for more efficient estimates (i.e., with lower variances), we can see clearly here that having additional symptoms that meet the assumptions of verbal autopsy studies will decrease the variance, but not affect the bias, of our estimates of cause-specific mortality.

Second, when the number of symptoms is large, direct tabulation can produce an extremely sparse matrix for $P(\boldsymbol{S})$ and $P(\boldsymbol{S}|D)$. For example, our data from China introduced in Section 2.6 have 56 symptoms, and so we would need to sort the $n = 1,411$ observations collected from the population into 2^{56} categories, which number more than 72 quadrillion. Reliable estimation by direct tabulation in this case is obviously infeasible. In practice, we only need to keep the symptom profiles that actually appear in both the hospital and population data sets, but even this can be sparsely populated. We thus develop an easy computational solution to this problem in the Appendix based on a varient of discrete kernel smoothing, which involves using random subsets of symptoms, solving (2.7) for each, and averaging. The difference here is that unlike the usual applications of kernel smoothing, which reduce variance at the expense of some bias, our procedure actually reduces both bias and variance here.

Third, the key statistical assumption of the method connecting the two samples is that $P(\boldsymbol{S}|D) = P^h(\boldsymbol{S}|D)$. If this expression holds in sample, then our method (and

indeed every subset calculation) will yield the true $P(D)$ population proportions exactly, regardless of the degree of sparseness. If the assumption instead holds only in the population from which the observed data are drawn, then our approach will yield statistically consistent estimates of the population density $P(D)$. If, in addition, subset sizes are small enough, then we find through simulation that our estimates are approximately unbiased.

Substantively, this key assumption would fail for example for symptoms that doctors make relatives more aware of in the hospital; following standard advice for writing survey questions simply and concretely can eliminate many of these issues. Another way this assumption can be violated would be if hospitals keep patients alive for certain diseases longer than they would be kept alive in the community, and as a result they experience different symptoms. In these examples, and others, an advantage of our approach, compared to approaches which model $P(D|S)$, is that researchers have the freedom to drop symptoms that would seem to severely violate the assumption.

Fourth, a reasonable question is whether expert knowledge from physicians or others could somehow be used to improve our estimation technique. This is indeed possible, via a Bayesian extension of our approach that we have also implemented. However, in experimenting with our methods with verbal autopsy researchers, we found few sufficiently confident of the information available to them from physicians and others that they would be willing to add Bayesian priors to the method described here. We thus do not develop our full Bayesian method here, but we note that if accurate prior information does exist in some application and were used, it would improve our estimates (see also Sibai et al. 2001).

Finally, the new approach represents a major change in perspective in the verbal autopsy field. The essential goal of the existing approach is to marshal the best methods to use $\boldsymbol{S}$ to predict D. The thought is that if we can only nail down the "correct" symptoms, and use them to generate predictions with high sensitivity and

specificity, we can get the right answer. There are corrections for when this fails, of course, but the conceptual perspective involves developing a *proxy* for D. That proxy can be well chosen symptoms or symptom profiles, or a particular aggregation of profiles as $\hat{D}$. The existing literature does not seem to offer methods for highly accurate predictions of D, even before we account for the difficulties in ascertaining the success of classifiers (Hand, 2006). Our alternative approach would also work well if symptoms or symptom profiles are chosen well enough to provide accurate predictions of D, but accurate predictions are unnecessary. In fact, choosing symptoms with higher sensitivity and specificity would not reduce bias in our approach, but in the existing approach they are required for unbiasedness except for lucky mathematical coincidences.

Instead of serving as proxies, symptoms in the new approach are only meant to be observable *implications* of D, and any subset of implications are fine. They need not be biological assays or in some way fundamental to the definition of the disease or injury or an exhaustive list. Symptoms need to occur with particular patterns more for some causes of death than others, but bigger differences do not help reduce bias (although they may reduce the variance). The key assumption of our approach is $P(\boldsymbol{S}|D) = P^h(\boldsymbol{S}|D)$. Since $\boldsymbol{S}$ is entirely separable into individual binary variables, we are at liberty to choose symptoms in order to make this assumption more likely to hold. The only other criteria for choosing symptoms, then, is the usual rules for reducing measurement error in surveys, such as reliability, question ordering effects, question wording, and ensuring that different types of respondents interpret the same symptom questions in similar ways. Other previously used criteria, such as sensitivity, specificity, false positive or negative rates, or other measures of predictability, are not of as much relevance as criteria for choosing symptom questions.

2.8 Implications for Individual Classifiers

We now briefly discuss the implications of our work for classification of the cause of each individual death. As the same results would seem to have broader implications for the general problem of individual classification in a variety of applications, we generalize the discussion here but retain our notation with $\boldsymbol{S}$ referring to what is called in the classifier literature features or covariates and D denoting category labels.

As Hand (2006, page 7) emphasizes, "Intrinsic to the classical supervised classification paradigm is the assumption that the data in the design set are randomly drawn from the same distribution as the points to be classified in the future." In other words, individual classifiers make the assumption that the *joint* distribution of the data is the same in the unlabeled (community) set as in the labeled (hospital) set $P(S, D) = P^h(S, D)$, a highly restrictive and often unrealistic condition. If $P(D|S)$ fits exceptionally well (i.e., with near 100% sensitivity and specificity), then this common joint distribution assumption is not necessary, but classifiers rarely fit that well.

In verbal autopsy applications, assuming common joint distributions or nearly perfect predictors is almost always wrong. Hand (2006) gives many reasons why these assumptions are wrong as well in many other types of classification problems. We add to his list a revealing fact suggested by our results above: Because $P(\boldsymbol{S})$ and $P^h(\boldsymbol{S})$ are directly estimable from the unlabeled and labeled sets respectively, these features of the joint distribution can be directly compared and this one aspect of the common joint distribution assumption can be tested directly. Of course, the fact that this assumption can be tested also implies that this aspect of the common joint distribution assumption need not be made in the first place. In particular, we have shown above that we need not assume that $P(\boldsymbol{S}) = P^h(\boldsymbol{S})$ or $P(D) = P^j(D)$ when trying to estimate the aggregate proportions. We show here that these assumptions

are also unnecessary in individual classifications.

Thus, instead of assuming a common joint distribution between the labeled and unlabeled sets, we make the considerably less restrictive assumption that only the *conditional* distributions are the same: $P(S|D) = P^h(S|D)$. (As above, we get the needed joint distribution in the unlabeled set by multiplying this conditional distribution estimated from the labeled set by the marginal distribution $P(S)$ estimated directly from the unlabeled set.) Thus, to generalize our results to apply to individual classification, which requires an estimate of $P(D_\ell|\boldsymbol{S}_\ell = \boldsymbol{s}_\ell)$, we use Bayes theorem:

$$P(D_\ell|\boldsymbol{S}_\ell = \boldsymbol{s}_\ell) = \frac{P(\boldsymbol{S}_\ell = s_\ell|D_\ell = j)P(D_\ell = j)}{P(\boldsymbol{S}_\ell = \boldsymbol{s}_\ell)} \tag{2.8}$$

We propose to use this by taking $P(\boldsymbol{S}_\ell = s_\ell|D_\ell = j)$ from the labeled set, the estimated value of $P(D_\ell = j)$ from the procedure described in Section 2.5, and $P(\boldsymbol{S}_i = \boldsymbol{s}_i)$ directly estimated nonparametrically from the unlabeled set, also as in Section 2.5. As with our procedure, we use subsets of $\boldsymbol{S}$ and average different estimates of $P(D_\ell|\boldsymbol{S}_i = \boldsymbol{s}_i)$, although this time the averaging is via committee methods since each subset implies a different model (with the result is constrained so that the individual classifications aggregate to the $\hat{P}(D)$ estimate). Each of these lower dimensional subsets (labled "sub") also imply easier-to-satisfy assumptions than the full conditional relationship, $P(\boldsymbol{S}_{\text{sub}}|D) = P^h(\boldsymbol{S}_{\text{sub}}|D)$.

We illustrate the power of these results with a simple simulation. For simplicity, we assume that features are independent conditional on the category labels in the labeled set, $P^h(\boldsymbol{S} = s|D) = \prod_{k=1}^{K} P(S_k = s_k|D)$, which is empirically reasonable except for heterogeneous residual categories. We then simulate data, with 5 (disease) categories, 20 (symptom) features, and 3000 observations in the labeled (hospital) and unlabeled (community) sets. We generate the data so they have very different marginal distributions for $P(\boldsymbol{S})$ and $P(D)$. Figure 2.3 gives these marginal distributions, plotting the unlabled set values horizontally and labeled set vertically; note

that few points are near the 45 degree line. These data are generated to violate the common joint distribution assumptions of all existing standard classifiers, but still meet the less restrictive conditional distribution assumption.

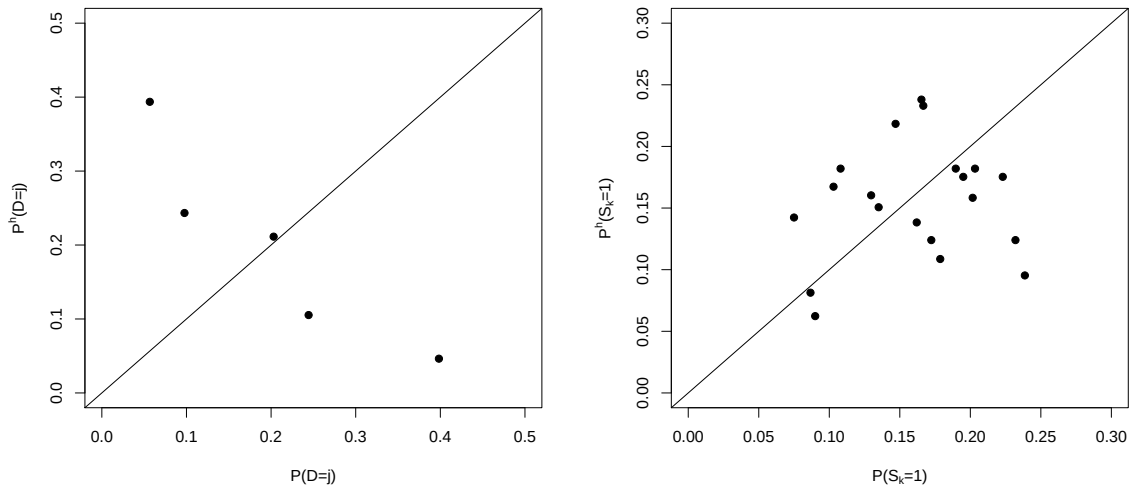

Figure 2.3: Simulated data. For both the proportion of observations in each category (in the left panel) and the proportion with each feature present (in the right panel), the labeled set is very different from the unlabeled target population of interest. These data would violate the assumptions underlying most existing classifiers.

We then run a standard support vector machine classifier (Chang and Lin, 2001) on the simulated data, which classifies only 40.5% of the observations correctly. In contrast, our simple nonparametric alternative classifies 59.8% of the same observations correctly. The key advantage here is coming from the adjustment of the marginals to fit $\hat{P}(D)$ in the "unlabeled" set. We can see this by viewing the aggregate results. These appear in Figure 2.4, with the truth plotted horizontally and estimates vertically. Note that our estimates (plotted with black disks) are much closer to the 45 degree line for every true value than the SVM estimates (plotted with

open circles).

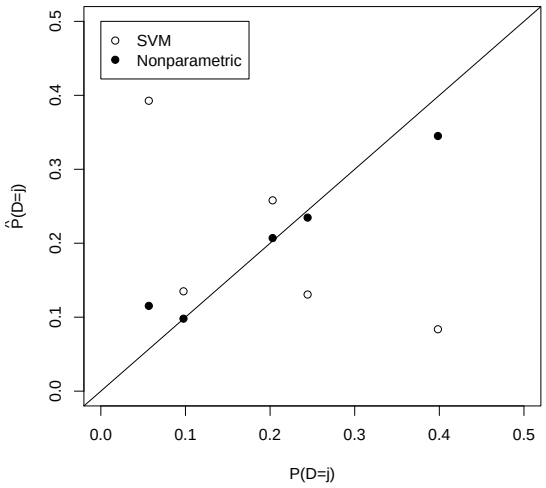

Figure 2.4: Individual-Level Classification by Support Vector Machine (open circles) and our improved Nonparametric Alternative (closed disks). Despite the differences between the labeled and unlabeled sets in Figure 2.3, our approach generates better aggregate results than the standard support vector machine classifier.

This section illustrates only the general implications of our strategy for individual classification. It should be straightforward to extend these results to provide a simple but powerful correction to any existing classifier, as well a more complete nonparametric classifier.

2.9 Concluding Remarks

By reducing the assumptions necessary for valid inference and making it possible to model all diseases simultaneously, the methods introduced here make it possible to extract considerably more information from verbal autopsy data, and as a result can

produce more accurate estimates of cause-specific mortality rates. Since our approach makes physician reviews, expert algorithms, and parametric statistical models unnecessary, it costs considerably less to implement and is easier to replicate in different settings and by different researchers. The resulting increased accuracy of our relatively automated statistical approach, compared to existing methods which require many more ad hoc human judgments, is consistent with a wide array of research in other fields (Dawes, Faust and Meehl, 1989)..

Even with the approach offered here, many issues remain. For example, to estimate the distribution of death by age, sex, or condition with our methods requires separate samples for each group. To save money and time, the methods developed here could also be extended to allow covariates, which would enable these group-specific effects to be estimated simultaneously from the same sample. A Bayesian approach could also be applied to borrow strength across these areas. A formal approach to choosing the smoothing parameter (the number of symptoms per subset) would be useful as well. In addition, scholars still need to work on reducing errors in eliciting symptom data from caregivers and validating the cause of death. Progress is needed on procedures for classifying causes of death and statistical procedures to correct for the remaining misclassifications, and on question wording, recall bias, question ordering effects, respondent selection, and interviewer training for symptom data. Crucial issues also remain in choosing a source of validation data for each study similar enough to the target population so that the necessary assumptions hold, and in developing procedures that can more effectively extrapolate assumptions from hospital to population via appropriate hospital subpopulations, data collection from community hospitals, or medical records for a sample of deaths in the target population.

2.10 Appendix: Estimation Methods

We now describe the details of our estimation strategy. Instead of trying to use all 2^K symptoms simultaneously, which will typically be infeasible given commonly used sample sizes, we recognize that only full rank subsets larger than J with sufficient data are required. We thus sample many subsets of symptoms, estimate $P(D)$ in each, and average the results (or if prior information is available we use a weighted average). To choose subsets, we could draw directly from the 2^K symptom profiles, but instead use the convenient approach of randomly drawing B ($B < K$) symptoms, which we index as $I(B)$, and use the resulting symptom sub-profile. This procedure is mathematically equivalent to imposing a version of kernel smoothing on an otherwise highly sparse estimation task. (More advanced versions of kernel smoothing might improve these estimates further.)

We estimate $P(\mathbf{S}_{I(B)})$ using the population data, and $P(\mathbf{S}_{I(B)} \mid D)$ using the hospital data. Denote $Y = P(\mathbf{S}_{I(B)})$ and $X = P(\mathbf{S}_{I(B)} \mid D)$, where Y is of length n, X is $n \times J$, and n is the subset of the 2^B symptom profiles that we observe. We obtain $P(D) \equiv \hat{\beta}$ by regressing Y on X under the constraint that elements of $\hat{\beta}$ fall on the simplex. The subset size B should be chosen to be large enough to reduce estimation variance (and so that the number of observed symptom profiles among the 2^B possible profiles is larger than J) and small enough to avoid the bias that would be incurred from sparse counts used to estimate of elements of $P(\boldsymbol{S}_{I(B)}|D)$. We handle missing data by deleting incomplete observations within each subset (another possibility would be model-based imputation). Although cross-validation can generate optimal choices for B, we find estimates of $P(D)$ to be relatively robust to choices of B within a reasonable range. (When choosing B via cross-validation from the hospital data, we use random subsets to separate this decision from the assumption that $P(\boldsymbol{S}|D) = P^h(\boldsymbol{S}|D)$.) We have experimented with nonlinear optimization

procedures to estimate $P(D)$ directly, but it tends to be sensitive to starting values when J is large. As an alternative, we developed the following estimation procedure, which tends to be much faster, more reliable, and accurate in practice.

We repeat the following two steps for each different subset of symptoms and then average the results. The two steps involve reparameterization, to ensure $\sum \beta_j = 1$, and stepwise deletion, to ensure $\beta_j > 0$.

1. To reparameterize, we follow this algorithm:

   (a) To impose a fixed value for some cause of death, $\sum \beta_j = c$, rewrite the constraint as $C\beta = 1$, where C is a J-row vector of $\frac{1}{c}$. When none of the elements of β are known a priori, $c = 1$. When we know some elements β_i, such as from another data source, the constraint on the rest of β changes to $\sum_{j \neq i} \beta_j = c = 1 - \beta_i$.

   (b) Construct a $J - 1 \times J$ matrix A of rank $J - 1$ whose rows are mutually orthogonal and also orthogonal to C, and so $CA^\top = 0$ and $AA^\top = I_{J-1}$. A Gram-Schmidt orthogonalization gives us a row-orthogonal matrix G whose first row is C, and the rest is A.

   (c) Rewrite the regressor as $X = ZA + WC$, where Z is $n \times J - 1$, W is $n \times 1$, and $(W, Z)G = X$. Under the constraint $C\beta = 1$, we have $Y = X\beta = ZA\beta + WC\beta = Z\gamma + W$, where $\gamma = A\beta$, and γ is a $J - 1$ vector.

   (d) Obtain the least square estimate $\hat{\gamma} = (Z^\top Z)^{-1} Z^\top (Y - W)$.

   (e) The equality constrained β is then $\hat{\beta} = G^{-1}\gamma^*$, where $G = (C, A)$, a $J \times J$ row-orthogonal matrix derived above,and $\gamma^* = (1, \hat{\gamma})$. This ensures that $C\hat{\beta} = 1$. Moreover, $\mathrm{Cov}(\hat{\beta}) = G^{-1}\mathrm{Cov}(\gamma^*)(G^\top)^{-1}$ (Thisted, 1988).

2. Then for Stepwise deletion:

(a) To impose nonnegativity, find the $\hat{\beta}_j < 0$ whose associated t-value is the biggest in absolute value among all $\hat{\beta} < 0$.

(b) Remove the j^{th} column of the regressor X, and go to the *reparameterization* step again to obtain $\hat{\beta}$ with the jth element coerced to zero.

Alternatively, we can view the estiamtion of β to be a constrained optimization problem and use the dual method to solve the strictly convex quadratic programs(citations..) Finally, our estimate of $P(D)$ can be obtained by averaging over the estimates based on each subset of symptoms. The associated standard error can be estimated by bootstrapping over the entire algorithm. Subsetting is required because of the size of the problem, but because $\boldsymbol{S}$ can be subdivided and our existing assumption $P(\boldsymbol{S}|D) = P^h(\boldsymbol{S}|D)$ implies $P(\boldsymbol{S}_{I(B)}|D) = P^h(S_{I(B)}|D)$ in each subset, no bias is introduced. In addition, although the procedure is statistically consistent (i.e., as $n \to \infty$ with K fixed) the procedure is approximately unbiased only when the elements of $P(\boldsymbol{S}|D)$ are reasonably well estimated; subsetting (serving as a version of kernel smoothing) has the advantage of increasing the density of information about the cells of this matrix, thus making the estimator approximately unbiased for a much smaller and reasonably sized sample. We find through extensive simulations that this procedure is approximately unbiased, and robust except in very small sample sizes.

Chapter 3

Understanding Complex Legislative and Judicial Behavior via Hierarchical Ideal Point Estimation

3.1 Introduction

In political science, researchers are often interested in understanding political actors' decision-making behavior. For example, why do the judges on the Supreme Court vote the way they do? And why do house members support some bills and reject others? Theorists of legislative and judicial behavior posit that political actors hold certain policy preferences or ideological values and such preferences underpin their voting behavior. However, there is usually no explicit data about political actors' preference. Instead, researchers seek to derive such information from alternative resources such as recorded vote data, speeches of political actors, or newspaper editorials.

In this paper, we focus on methods of ideal point estimation that measure political preferences through recorded vote data. A hierarchical statistical framework for ideal point estimation is introduced. Under this framework, researchers can model correlated voting behavior among groups of individuals and each individual's decisions on related issues. In particular, the hierarchical structure is implemented to allow the elucidation of the characteristics of the decision makers and of the pending bills/cases. Hence, this framework entails a substantively intuitive and statistically coherent approach to address important issues such as the multidimensional nature of political ideology, party or interest group influence, period effect and strategic voting.

The rest of this paper is organized as follows. In Section 3.2, we briefly review existing ideal point estimation research, and discuss how the complexity in voting behavior could falsify the commonly adopted assumptions regarding independent voting. Section 3 introduces our model and an schematic illustration of modeling correlated voting behavior through hierarchical structures. Section 4 highlights model estimation. Section 5 presents the results of simulation studies to assess model performance. Section 6 follows with two empirical examples. One example analyzes the legislative behaviors of the 109th US house of representatives and the other analyzes the judicial behaviors of the US supreme court justices (1919-1996). Finally, Section 7 concludes the paper with a discussion about the potential of this framework.

3.2 Traditional Ideal Point Estimation and Correlated Voting Behavior

In political science research, the quantitative measurement of political preference is typically done from ideal points (Epstein & Mershon, 1996; Poole & Rosenthal, 1997; Segal & Spaeth, 1997; Jackman, 2000; Longdregan, 2000; Martin & Quinn,

2002; Clinton et.al., 2004; Poole, 2005). The estimation of ideal points is based on a theoretical construct of ideological space which represents a liberal-conservative continuum (Poole & Rosenthal, 1997). The main goal of ideal point estimation is to uncover the position of each legislator in the ideological space based on observed vote records.

Suppose there are I political actors making decisions on J different items. The items can be bills discussed in the Congress or cases in the court. The decisions are recorded in a data matrix $\{y_{ij}, i = 1, \dots I, j = 1, \dots, J\}$. When we observe a "Yea" vote on the jth item by the ith legislator, $y_{ij} = 1$; when we observe a "Nay" vote, $y_{ij} = 0$. Following Clinton, et.al. (2004), a unidimensional ideal point estimation model is based on a latent score t_{ij} with the following form

$$t_{ij} = a_j \theta_i - b_j + \epsilon_{ij}, \tag{3.1}$$

and

$$y_{ij} = \begin{cases} 1 & \text{if} \quad t_{ij} \geq 0 \\ 0 & \text{if} \quad t_{ij} < 0 \end{cases}$$

where θ_i is the ideal point of the ith individual, b_j measures how difficult it is for an individual to agree with the jth item, a_j measures the direction and sensitivity of the jth item in distinguishing individuals' ideal points, and ϵ_{ij} is the identically and independently distributed error term.

In recent years, more attention has been paid to complex ideological structures and voting behavior. For examples, to what extent does partisanship influence legislators' voting behaviors? Is the unidimensional ideological space sufficient to summarize the variation of political preference? Other questions of interest include the temporal changes of legislators' political preference, committee voting and strategic voting. Much of the work has been done based on an ad hoc analysis of ideal point estimates from models, such as equation (3.1), and they often suffer from mis-specification

problems. There have been several attempts to extend the classical model, such as multidimensional ideological estimation (e.g., Jackman, 2001; Rivers, 2003) and dynamic ideal point estimation (Martin and Quinn, 2002). But those models are not only computationally challenging due to rapidly increasing number of parameters to be estimated, they are also difficult to interpret since no information about the characteristics of the legislators and the contents of bills and cases that are presented to the legislators is incorporated into the estimation.

Under the classical ideal point model (1), there are two assumptions of independence: given item j, every individual votes independently; given ith individual's ideal point, he/she votes independently across all items. The first independence assumption could be violated when voters are influenced by peers, for example, when party members are influenced to vote toward the party policy line regardless of their ideological values. The second independent assumption is termed *Local Item Independence* (LII) in psychometric literature. One situation in which it fails is when the unidimensional model is insufficient. For example, a justice can be socially liberal, but conservative concerning economic issues. Another source of local item dependence is related to temporal changes in political preference. In different time periods, legislators could vote differently in responding to the changes in political institutions. For example, Lu and McFarland (2007) found that in the US house of representatives, there are significant period patterns in voting behavior of the congressmen under unified democratic, divided and unified republican government in the last ten US congresses. When the independence assumptions are violated, the parameter estimates θ_is and a_j, b_js conditional on those independencies will be biased and inefficient. There are many references in psychometrics discussing this issue (see Sireci et. al., 1991, Wainer and Thissen, 1996 and Yen, 1993).

In this paper, we generalize equation (1) to allow the characteristics of the political actors and of the cases/bills as well as the context in which the votes are cast to be

modeled in ideal point estimation. Specifically, the generalization takes the following form

$$t_{ij} = a_j(\theta_{ik} + \delta_m) - b_{lj} + \epsilon_{ij}$$

where we allow the ideal point θ_i to vary across subgroups of the bills/cases of different contents and the item parameter b_j to vary across subgroups of individuals of different characteristics. δ_m is an additional term that capture transitory drifts across subsets of votes cast by subgroup of individuals. In other words, we model the "covariates" into ideal point estimation. Naturally this model takes care of the deviation of the assumptions of independence by introducing random effect terms θ_{ik}, b_{lj} and δ_m, and allowing them to interact with individuals and cases. Hence the different dependence structures in ideal point estimation can be conveniently modeled. A detailed discussion of this model will be presented in Sections 3 and 4.

3.3 Model Complex Dependent Structure

To illustrate our model, we first introduce a set of definitions to denote different types of dependent structures in the recorded vote data. The chart below illustrates these definitions in a political voting context where there are 10 legislators casting votes on 12 bills. The party affiliations of the legislators are labeled $\mathcal{D}$, $\mathcal{R}$, or blank if the legislator does not belong to either party. Moreover, we assume that the contents of the 12 bills can be classified into three different issue areas: economic activities ($\mathcal{EA}$), civil liberties issues ($\mathcal{CL}$) and political issues ($\mathcal{PI}$). A vote cast by a legislator is indicated by an "$\cdot$". The symbols $\star$, $*$ and \$ represent the three different voting blocks.

	J	1	2	3	4	5	6	7	8	9	10	11	12
I		$\mathcal{EA}$	$\mathcal{EA}$	$\mathcal{EA}$	$\mathcal{EA}$	$\mathcal{CL}$	$\mathcal{CL}$	$\mathcal{CL}$	$\mathcal{CL}$	$\mathcal{PI}$	$\mathcal{PI}$	$\mathcal{PI}$	$\mathcal{PI}$
1	$\mathcal{D}$	·	·	·⋆	·⋆	·⋆	·	·	·	·	·	·	·
2	$\mathcal{D}$	·	·	·⋆	·⋆	·⋆	·	·	·	·	·	·	·
3	$\mathcal{D}$	·	·	·⋆	·⋆	·⋆	·	·	·	·	·	·	·
4	$\mathcal{D}$	·	·	·	·	·	·	·	·	·	·∗	·∗	·∗
5	$\mathcal{R}$	·	·	·	·	·	·	·	·	·	·∗	·∗	·∗
6	$\mathcal{R}$	·	·	·	·	·	·	·	·	·	·∗	·∗	·∗
7	$\mathcal{R}$	·	·	·	·	·	·	·	·	·	·∗	·∗	·∗
8	$\mathcal{R}$	·	·	·$	·$	·	·	·	·	·	·	·	·
9		·	·	·$	·$	·	·	·	·	·	·	·	·
10		·	·	·$	·$	·	·	·	·	·	·	·	·

We use the term *allyset* to denote a group of individuals who typically vote together; the term *voteset* to denote a cluster of items to which each individual's decisions are correlated; and the term *tactset* to indicate a block of correlated decisions made by a group of individuals on a cluster of items.

- *allyset*: The hierarchical structure among individuals is defined by *allysets*. An *allyset* consists of individuals who tend to influence each other when they vote. For example, individuals who belong to party $\mathcal{D}$ can be considered as an *allyset*, while individuals labeled $\mathcal{R}$ belong to another *allyset*. *Allysets* are flexible constructs which can be determined by the characteristics of the political actors. Furthermore, not all individuals need to be included in an *allyset*; individual voters can coexist with *allysets*.

- *voteset*: The hierarchical structure among items is delineated by *votesets*. Specifically this term denotes a cluster of bills/cases to which each individual's deci-

sions are correlated; A *voteset* can be determined by the characteristics of the items such as issue areas or time periods. For example, bills 1-4 belong to same *voteset* $\mathcal{EA}$ because they all concern economic activities, while bills 5-8 and 9-12 belong to two other *votesets*.

- *tactset*: *tactsets* are flexible constructs. They are blocks (or collections) of correlated decisions that are made by a subgroup of individuals on a selection of bills. In the chart above, the components of a *tactset* are indicated by the additional symbols $\star$, *, or \$. In general, a *tactset* consists of votes that are affected by temporary political coalitions. An example of *tactset* are the decisions made by individuals 8, 9 and 10 on bills 3 and 4 driven by shared economic incentives. *Tactsets* can also represent strategic votes. For example, individuals 1, 2 and 3 can vote strategically on bills 3-5 in trade for preferred outcomes of individuals 4 to 7 voting on bills 10-12.

Conditional on *allysets*, *votesets* and *tactsets*, we can write a hierarchical ideal point estimation model as follows,

$$P\left(\, y_{ij} = 1 |\, d(j) = k, p(i) = l, r(i,j) = m\right) = \Phi(a_j(\theta_{ik} + \delta_m) - b_{lj}), \tag{3.2}$$

with the corresponding latent function,

$$t_{ij} = a_j(\theta_{ik} + \delta_m) - b_{lj} + \epsilon_{ij} \tag{3.3}$$

where ϵ_{ij} is a standard normal error. *Votesets* are indexed by $d(j)$; if question j belongs to *voteset* k, $d(j) = k$. When making decisions, the model assumes each individual has a unique and *voteset*-specific ideal point, θ_{ik}. *Allysets* are indexed by $p(i)$; if individual i belongs to *allyset* l, $p(i) = l$. The term b_{lj} then measures how likely it is for members of the lth *allyset* to agree with the jth item. Hence we can view b_{lj} as the lth *allyset*'s policy position for the jth item. Lastly, the effect

of *tactset*–transitory coalition is denoted by $\delta_{r(i,j)}$ with *tactset* indexed by $r(i,j)$. If individual i and item j belong to *tactset* m, $r(i,j) = m$. For individuals and items that do not belong to any of the *tactsets*, $r(i,j) = 0$ and $\delta_{r(i,j)} = 0$.

We model $\theta_{ik}, k = 1, \ldots, K$ as random effects following Multivariate Normal Distribution with mean $\mathbf{0}$ and variance-covariance matrix Σ_K. K is the number of *votesets*.

$$(\theta_{i1}, \ldots, \theta_{iK})^t \sim \text{MVN}\,(\,0,\, \Sigma_K)\,, \quad i = 1, \ldots, I,$$

where the mean equal to $\mathbf{0}$ is for model identification purpose and the diagonal elements of Σ_K are the variance of the *votesets*, σ_k^2. The off-diagonal terms are the covariance between two *votesets*, $\sigma_{kk'}$. We model each distinct value of b_{lj} as independent random effects following Multivariate Normal Distribution with mean $\mu^{\mathbf{b}}$ and variance Ψ_G where G is the number of *allysets*.

$$(b_{1j}, \ldots, b_{Gj})^t \sim \text{MVN}\left(\mu^{\mathbf{b}}, \Psi_G\right), \quad j = 1, \ldots, J.$$

Last, we model δ_m through

$$\delta_m \sim \text{N}(\mu_\delta, \sigma_\delta^2),\ m = 1, \ldots, M.$$

We complete the specification of model (3.3) into a larger Bayesian hierarchical framework by treating item parameter a_j as random effects. Specifically we assume:

$$a_j \ \sim \ \text{N}(\mu_a, \sigma_a^2)$$

Compared to the traditional ideal point estimation model, now we can model various dependencies introduced by the *votesets* and *allysets*. For example, in model (1) the within-person correlation and within-item correlation of the latent scores are assumed to be constant.

$$\text{cor}\,(t_{ij}, t_{ij'}) = \frac{\mu_a^2 \sigma_\theta^2}{V(t_{ij})},\ j \neq j', \qquad \text{cor}\,(t_{i'j}, t_{ij}) = \frac{\sigma_b^2}{V(t_{ij})},\ i \neq i'.$$

where the variance of t_{ij} is a constant, $V(t_{ij}) = 1 + \sigma_b^2 + (\mu_a^2 + \sigma_a^2)\sigma_\theta^2$. However, when there are *votesets*, $\text{cor}(t_{ij}, t_{ij'})$ depends on whether items j and j' belong to the same *voteset* or not. Under the hierarchial model (3), assuming no *allysets* to simplify the situation:

$$\text{cor}\,(t_{ij}, t_{ij'}) = \begin{cases} \frac{\mu_a^2\sigma_k^2}{(\sigma_a^2+\mu_a^2)\sigma_k^2+\sigma_b^2+1}, & \text{if } d(j) = d(j') = k \\ \frac{\mu_a^2\sigma_{kk'}}{\sqrt{(\sigma_a^2+\mu_a^2)\sigma_k^2+\sigma_b^2+1}\sqrt{(\sigma_a^2+\mu_a^2)\sigma_{k'}^2+\sigma_b^2+1}}, & \text{if } d(j) = k, d(j') = k', k \neq k' \end{cases}$$

We note that for ith person, the correlation between his/her responses to two items j and j' that belong to the same *voteset* k is proportional to the variance of that *voteset*, σ_k^2. Hence the larger the variance, the more significant the *voteset* effect. On the other hand, if the two items do not belong to the same *voteset*, the correlation of the corresponding latent scores depends on $\sigma_{kk'}$, the covariance between the two *votesets*. Similarly, where there are *allysets*, we can capture the dependencies within and across *allysets*.

In model (3), the covariance matrix Σ_k reveals the dependence structure of the *votesets*. We can derive the correlation between any two *votesets*, $\rho_{kk'} = \sigma_{kk'}/(\sigma_k\sigma_{k'})$. The value $\rho_{kk'} = -1$ indicates maximal deviation of individual's ideological values in the two *votesets* k and k'; the same individual who is liberally minded when making decisions in *voteset* k can be conservatively minded when making decisions in *voteset* k'. When $\rho_{kk'} = 0$, the individual makes a decision in *voteset* k independently from decisions in *voteset* k'. In this situation, knowing the liberalism/conservativeness of an individual i in *voteset* k does not help us to infer his/her ideological stand in the other *voteset* k'. On the other hand, if the two *votesets* are highly positively correlated, one could consider combining them.

Similarly, the covariance matrix Ψ_G models how *allysets* affect the vote outcomes and the interactions among them. A larger variance of *allyset* l implies a more unified vote outcomes among the *allyset* members than the sincere vote outcomes based solely

on their ideal points, hence more significant group effect. The correlation $\rho_{ll'}$ between any two *allysets* measures the extent to which these two groups cooperate with or obstruct each other. When there is little difference between the policy positions induced by two *allysets*, $\rho_{ll'} \to 1$. On the other hand, a small correlation or even negative correlation implies that a fair number of votes are influenced by the different party agendas.

In this hierarchical framework, *allysets*, *votesets* and *tactsets* are defined by the contents of the bills or the cases, the characteristics of the voters as well as the context in which they vote. Hence researchers are empowered to directly incorporate their substantive knowledge about legislative and judicial behaviors into statistical modeling. Furthermore, such definitions of *allysets*, *votesets* and *tactsets* make them very malleable structures and they offer researchers opportunities to test alternative theories of voting. In the following sections, we will demonstrate models with only *allysets* or *votesets* through simulations and real data application. We will discuss the cases of *tactsets* and any combinations of three dependent structures in Section 3.7 for future applications. Therefore, in the rest of the paper, we focus on the following reduced model

$$t_{ij} = a_j \theta_{id(j)} - b_{p(i)j} + \epsilon_{ij} \tag{3.4}$$

3.4 Hierarchical Ideal Point Estimation

3.4.1 The identification of the model

Without constraints, model (3.4) is in general non-identifiable. Bafumi et.al. (2005) pointed out two sources of identification problems:

$$\text{additive aliasing:} \qquad a_j\theta_{ik} - b_{lj} = a_j(\theta_{ik} + c_0) - (b_{lj} + c_0)$$
$$\text{multiplicative aliasing:} \qquad a_j\theta_{ik} - b_{lj} = \left(\frac{a_j}{d_0}\right)(\theta_{ik}d_0) - b_{lj}$$

To solve the additive aliasing problem, we constrain the ideal points to be mean 0.

$$\mathrm{E}(\theta_{ik}) = 0.$$

Since $a_j \in (-\infty, \infty)$, one aspect of the multiplicative aliasing problem is reflection invariance, i.e., $a_j\theta_{ik} = (-)a_j(-)\theta_{ik}$. To prevent this problem, a constraint can be put on the rank orders of at least two θ_{ik} values of each *voteset*. In a political voting context, this can be easily done by identifying a subset of legislators that are quoted as being highly conservative or highly liberal in each *voteset*. To completely solve the multiplicative aliasing problem, constraints also need to be imposed on the variance of either *votesets* θ_{ik} or a_j. As will be shown in Section 3.4.2, we handle this problem by applying an informative prior distribution to θ_{ik} such that the posterior distribution is not subject to multiplicative aliasing. Essentially the political interpretation of the ideal points, *allysets* effects and *votesets* effects are not affected by these constraints since they are invariant to the scale changes.

3.4.2 Estimation of the proposed models

The large number of parameters in model (3.4) invites a fully specified Bayesian framework. Advances in Bayesian estimation such as Gibbs sampler and Markov Chain Monte Carlo make the posterior simulation of large number of parameters tractable and efficient.

To minimize the impact of prior specification on the posterior estimation, we apply

noninformative Jeffrey's prior distributions to the hyper-parameters.

$$\begin{aligned}\left(\mu^{\mathbf{b}}, \Psi_G\right) &\sim \mid \Psi_G \mid^{-(L+1)/2}, \\ (\mu_a, \sigma_a^2) &\sim \sigma_a^{-2}.\end{aligned}$$

Then we choose the prior distribution for the *voteset* effects θ_{ik} in order to solve the multiplicative aliasing problem as follows:

$$(\theta_{i1}, \ldots, \theta_{iK})^t \sim \text{MVN}\left(0,\, I_K\right), \quad i = 1, \ldots, I,$$

where I_K is a K-dimensional identity matrix. Then the conditional posterior distribution of θ_{ik} given the latent score t_{ij} and the posterior draws of $a_j, b_{lj}, p(i) = l$, as follows,

$$\begin{aligned}f(\theta_{ik} \mid t_{ij}, a_j, b_{lj}) \;\propto\; \exp\Bigg\{-\frac{1}{2}\Bigg[&(\sum_j a_j^2 + 1)\theta_{ik}^2 - 2\sum_j (t_{ij} - b_{lj})a_j\theta_{ik} \\ &+ \sum_j (t_{ij} - b_{lj})^2\Bigg]\Bigg\} \qquad (3.5)\end{aligned}$$

where $\sum_j$ is a summation over $j \in voteset_k$.

Under the above prior specifications, the posterior estimation of model (3.4) is fairly straightforward via data augmentation (Tanner and Wong, 1988) and Gibbs sampler. In particular, at each step of Gibbs sampler, the conditional posterior distributions have closed forms. For details see Lu & Wang (2008) (Software manual).

3.4.3 Assessing the model fit

As the entire estimation is carried through a Bayesian framework, one can access the model fitting through the posterior predictive checks, for example, calculating the latent continuous residual or the Mean Absolute Predictive Error. Moreover, researchers can test hypotheses through the posterior p-value. For example, one can

test whether there is difference in the variances of two *allysets* by comparing their posterior draws.

The formation of *allysets*, *votesets* and *tactsets* are all subjective. To choose the right formation, we use the Deviance Information Criterion (DIC) for model selection, which is a Bayesian model selection rule developed by Spiegelhalter et al. (2002). DIC is an extension to the Akaike information criterion and it takes the following form:

$$\text{DIC} = \bar{D} + p_D,$$

where $\bar{D}$, the posterior expectation of the deviance, serves as the Bayesian measure of model adequacy. It can be calculated directly from the MCMC chains. p_D is a penalty term that accounts for the complexity of the model. Due to the structure of the Bayesian hierarchical model and the shared prior distribution, the parameters are in general not independent. The value p_D in this case can be interpreted as "the effective number of parameters". In general, the smaller the value of DIC, the better the model fitting.

3.5 Simulation Studies

In this section, we present a set of simulation studies to assess the performance of our model in estimating *allyset* effects and *voteset* effects.

There are three different hypothetical scenarios from which simulated data sets are generated. In the first scenario it is assumed that 100 individuals vote on 100 items. Moreover, there are 4 *votesets*, each of which has 25 items; and 2 *allysets*, each of which has 50 individuals. The second scenario assumes 200 individuals voting on 200 items. As in the first scenario, there are 4 *votesets* of equal numbers of items and 2 *allysets* of equal numbers of individuals. Lastly the third scenario involves 400 individuals voting on 400 items. It has the same structure of *votesets* and *allysets*.

Moreover, in each of the scenarios, *voteset* 1 and *voteset* 2 are assumed to be correlated with correlation 0.5, and *voteset* 1 and *voteset* 3 are correlated with correlation -0.8. The rest of the *votesets* are assumed to be pair-wise independent. Hence, in each simulation, the *voteset* effects θ_{ik}, $k = 1, \ldots, 4$, are drawn from a multi-variate normal distribution as follows:

$$\begin{pmatrix} \theta_{i1} \\ \theta_{i2} \\ \theta_{i3} \\ \theta_{i4} \end{pmatrix} \sim \text{MVN} \left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.5 & -0.8 & 0 \\ 0.5 & 1 & 0 & 0 \\ -0.8 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \right).$$

The two *allysets* are correlated with correlation -0.5 and they have different mean and variance. Specifically, in each simulation, the *allyset* effects $b_{lj}, l = 1, 2$, are drawn from a bivariate normal distribution as follows:

$$\begin{pmatrix} b_{1j} \\ b_{2j} \end{pmatrix} \sim \text{BVN} \left(\begin{pmatrix} 1 \\ -1 \end{pmatrix}, \begin{pmatrix} 2 & -0.7071 \\ -0.7071 & 1 \end{pmatrix} \right).$$

To directly assess the performance of the proposed Bayesian model estimation, we simulate 100 datasets (due to the expansive computation of Bayesian method) for each scenario and fit the Bayesian model for each data set. In Table 3.1, the estimated expected values (est.) of the following parameters are reported: the posterior variance of the *voteset* effects θ_{ik} and the correlation between every two *votesets*, the posterior mean and variance of the *allyset* effects b_{lj} and the correlation between the two *allysets*, and the posterior mean and variance of item parameter a_j. The expected values of these parameter estimates are estimated by averaging the results based on the 100 simulations. The estimated standard errors (s.e.) of these parameter estimates are also reported. Table 3.1 indicates that the estimated entries for the covariance matrix of θ tend to be underestimated when sample size is small. This is due to the shrinkage

effect of Bayesian modeling. However, this effect decreases as sample size increases. Similarly, as sample size increases, the parameter estimates of the *allyset* effects and of a_js approach to their true values with higher precision.

Table 2 reports the mean square errors of the individual level *voteset* specific ideal point estimates, $\sum_{i=1}^{I}(\hat{\theta}_{ik} - \theta_{ik})^2/I$, the mean square errors of the item level *allyset* specific effects estimates, $\sum_{j=1}^{J}(\hat{b}_{lj} - b_{lj})^2/J$, and the mean square error of a_j, $\sum_{j=1}^{J}(\hat{a}_j - a_j)^2)/J$. As the number of questions and the number of individuals double, the mean square errors of these individual level and item level parameters are also halved. In general, these simulation studies show that the Bayesian estimation gives consistent results.

3.6 Applications to US Judicial and Legislative Behavior

In this section, we offer two examples of applying the hierarchical ideal point estimation model to analyze legislative and judicial behavior. In the first example, treating each party as an *allyset*, we examine to what extent party affiliations affect decisions made by the members of the 109th US congress. The second example explores the issue-specific preferences of the US Supreme Court justices in different issue areas defined by *votesets*.

To ensure the convergence of estimating a large number of parameters, all the results we present are based on multiple chains and past convergence diagnostics (Gelman and Rubin, 1992). Four Markov Chains are run in each example; each chain consists of 10,000 iterations after a burn-in period of 10,000 iterations, and only every 10th draw is kept in order to reduce the serial correlation of the Markov Chains.

In both data sets, especially in the Supreme Court data, there are many missing

Table 3.1: Parameter estimation of the simulated examples.

	Simulation 1	Simulation 2	Simulation 3
	$I = 100, J = 100$	$I = 200, J = 200$	$I = 400, J = 400$
parameter	est.(s.e.)	est.(s.e.)	est.(s.e.)
voteset			
$V(\theta_1) = 1$	0.876 (0.067)	0.931 (0.040)	0.975 (0.024)
$V(\theta_2) = 1$	0.869 (0.072)	0.938 (0.044)	0.969 (0.026)
$V(\theta_3) = 1$	0.874 (0.058)	0.939 (0.037)	0.965 (0.028)
$V(\theta_4) = 1$	0.875 (0.063)	0.937 (0.038)	0.965 (0.034)
$\rho_\theta(1,2) = 0.5$	0.438 (0.086)	0.478 (0.052)	0.486 (0.040)
$\rho_\theta(1,3) = -0.8$	-0.711 (0.042)	-0.758 (0.028)	-0.776 (0.019)
$\rho_\theta(1,4) = 0$	-0.000 (0.102)	-0.011 (0.061)	0.000 (0.050)
$\rho_\theta(2,3) = 0$	0.006 (0.092)	-0.006 (0.068)	-0.002 (0.049)
$\rho_\theta(2,4) = 0$	0.002 (0.102)	-0.005 (0.070)	-0.001 (0.047)
$\rho_\theta(3,4) = 0$	-0.013 (0.090)	0.001 (0.069)	0.001 (0.050)
allyset			
$E(b_1) = 1$	1.012 (0.137)	1.004 (0.097)	0.999 (0.075)
$E(b_2) = -1$	-1.002 (0.089)	-1.011 (0.065)	-0.994 (0.043)
$V(b_1) = 2$	2.159 (0.399)	2.032 (0.242)	2.058 (0.174)
$V(b_2) = 1$	1.058 (0.174)	1.014 (0.114)	1.011 (0.075)
$\rho_b(1,2) = -0.5$	-0.497 (0.081)	-0.501 (0.057)	-0.509 (0.040)
a			
$E(a) = 0$	-0.001 (0.136)	0.010 (0.100)	0.000 (0.073)
$V(a) = 2$	2.068 (0.437)	2.050 (0.239)	2.013 (0.159)

Table 3.2: Mean square errors of the individual level and item level parameters.

parameter	Simulation 1 $I = 100, J = 100$ Est.(s.e.)	Simulation 2 $I = 200, J = 200$ Est.(s.e.)	Simulation 3 $I = 400, J = 400$ Est.(s.e.)
Mse(θ_1)	0.128(0.028)	0.066(0.014)	0.032(0.005)
Mse(θ_2)	0.130(0.037)	0.062(0.012)	0.032(0.005)
Mse(θ_3)	0.127(0.034)	0.064(0.013)	0.032(0.006)
Mse(θ_4)	0.122(0.031)	0.063(0.011)	0.031(0.004)
Mse(b_1)	0.201(0.056)	0.112(0.030)	0.062(0.012)
Mse(b_2)	0.121(0.025)	0.068(0.010)	0.035(0.005)
Mse(a)	0.142(0.031)	0.068(0.011)	0.034(0.004)

values in the data matrices. To minimize the impact of these missing values on model fit, we treat them as missing at random.

3.6.1 Party effect in Congress

The roll call records in the US house of Representatives are known to be very polarized. For example, 48% of the bills voted during the term of the 109th Congress were unanimously voted by at least one party. Researchers have been studying the effect of party in roll call voting, specifically, whether the observed polarization is due to the sharp division of political preference between the Democrats and the Republicans, or is due to party pressure during the roll call processes. There have been some attempts to measure the party effect in roll call voting, for examples, by comparing the ideal points of the median legislator within each party (Schickler, 2000); by examining the behavioral changes in individual party switchers (Clinton, Jackman

and Rivers, 2004); or by modeling party induced effect (see e.g., party inducement model, Snyder and Groseclose, 2000; Clinton, et.al., 2004). However, none of these approaches directly address the problem that the existence of party effect violates the assumption of independent voting. Hence these models suffer from inaccurate model specifications and inability to quantify party effect.

In this section, we will explore the effect of party in the context of correlated voting behavior. To examine the extent to which parties shape individuals' voting behavior, we fit a hierarchical ideal point estimation model with two *allysets* that are defined by the Democratic party and the Republican party in the Congress. Our strategy is to estimate the ideal point of each House member assuming a party specific policy position, the *allyset* effect $b_{lj}, l = 1, 2$, for each roll call j. Meanwhile, we can obtain estimates of the true ideal point of each legislator, while controlling for party influence. For comparison purposes, we also fit standard ideal point model (the null model, model (3.1)) through a Bayesian framework based on the assumption of independent voting.

In this analysis, we take the roll call data of the 109th Congress compiled by Lewis and Poole (2008). Unanimous votes were excluded from the analysis since they do not offer any information in distinguishing the ideal points of the legislators. There are 1038 non-unanimous votes cast by 440 representatives of the 109th congress, in which 203 were Democrats and 237 were Republicans. Figure 3.1 summarizes the results based on these data.

The x-axes of the two graphs on the left plot the estimations of the ideal points of house members before and after controlling for the effects of the *allysets* (null model versus the party model). Each dot represents a member of the congress; solid dots indicate Democrats and circled dots indicate Republicans. To identify the model, we constrain the most liberal member of the house to have a negative valued ideal point and vice versa for the most conservative member of the house. Hence, the more

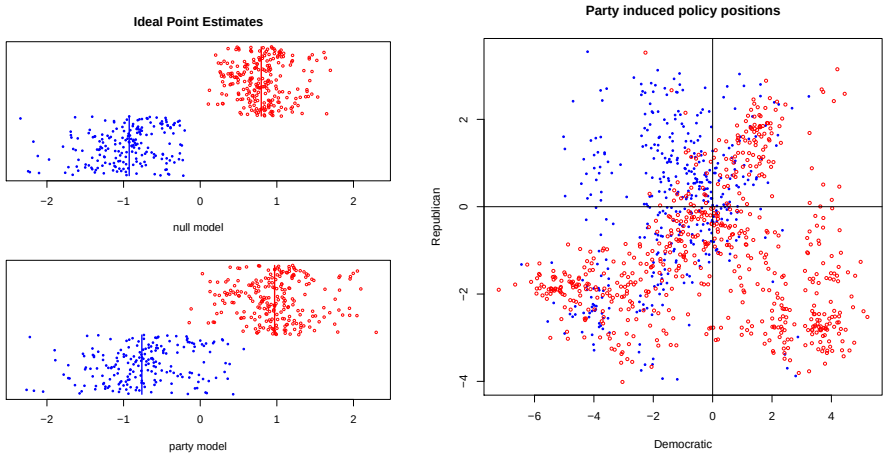

Figure 3.1: Ideal point estimates of members of US congress and party induced policy positions.

negative the ideal point, the more liberal the legislator is.

The most distinctive feature of these two graphs is that after controlling for party effect, there is significant overlap of the ideal points of the two parties. In contrast, the null model suggests absolute polarization. This shows that the observed dramatic separation between the Democrats and the Republicans (as modeled under the null model) is partially due to the party effect, rather than stark separations of the ideological preferences between the individuals of the two parties. Moreover, it is interesting to see that under the party model, the ideological positions of both moderates and extremists of each party extend further away from the party median positions (as indicated by the vertical lines), which implies that parties tend to influence the vote outcomes by squeezing both moderates and extremists closer to the party line (compared to the observed ones under the null model). On the other hand, we also observe that the median positions of the ideal points for both are still widely separated, and change little between the two models. Therefore, in general, the roll

call outcomes are congruent with most House members' ideological preferences even under significant party influence.

Furthermore, our model reveals that the variance of the *allyset* Democrat is 6.7 and the variance of *allyset* Republican is 2.8. This shows that when voting on each bill, there is a much higher correlation among the decisions made by Democratic party members than among the Republican party members. Hence the Democratic party members tend to have more unified votes than the Republican party given their ideal points. This finding is not surprising since being the minority party in the 109th Congress, the Democratic party might be expected to influence the vote results to maximize the party's power.

Moreover, the correlation between these two *allysets* is just 0.07, suggesting that the two parties obstruct each other on a large number of roll calls. In the right panel of Figure reffig:party, we plot the Democratic *allyset* effects b_{1j} against the Republican *allyset* effects b_{2j}. A bill with liberal content ($a_j < 0$) is labeled by a solid dot and a bill with conservative content ($a_j > 0$) is labeled by a circle. In this scatter plot, about 40% of the dots are located in the 1st and 4th quadrangles. This shows that both parties move the policy positions of those bills toward opposite directions for favorable vote outcomes. In particular, data points in the 1st quadrangle mostly consist of solid dots, suggesting that the Democratic party tends to make a sizable number of the bills with liberal contents more passable (with negative b_{1j} values) among their members while the Republican party does the opposite. A mirror image can be seen in the 4th quadrangle suggesting both parties manipulate a significant number of the conservatively contented roll calls as well. Further research is called for to investigate which types of roll calls are more subject to party manipulation.

Lastly, we present the Deviance Information Criterion (DIC) and Mean Absolute Predictive Errors (MAPE) of these two models (see Table 3.3. Not surprisingly, the party *allyset* model performs much better than the null model. To compare the

performance of non-informative prior, we also present the DIC and MAPE of the standard software package MCMCpack where informative priors are used (Martin & Quinn, 2008). It seems that the model is not sensitive to the different priors.

Table 3.3: Deviance Information Criterion and Mean Absolute Predictive Errors of 3 models.

Model	DIC	$\bar{D}$	p_D	MAPE
Null model	145248	143565.3	1682.7	0.167
Party model	140100.5	137850.3	2250.2	0.159
MCMCirt1d	145233.6	143550.7	1682.8	0.167

3.6.2 Estimating ideal points within different issue areas

Being the individuals with the highest judicial power in the US, the Supreme Court justices receive a lot of attention on their ideological values. Earlier substantive research has suggested that the decisions justices make could have come from different ideological dimensions. For examples, Schubert (1974) suggests that the ideological values of the justices can be summarized within two value systems–civil liberties and economic liberties. Spaeth (1979) suggests the justices support or oppose the three values of freedom, equality and New Deal economics. On the other hand, more recent research based on the court rulings records under chief justice Rehnquist found evidence of uni-dimensional court via pattern analysis (Sirovich, 2003). Recently, efforts have been made to develop multidimensional ideal point estimation (Jackman, 2001; Rivers, 2003, Poole, 2005). However, such models are all based on the assumption of orthogonal multidimensional space which suffers lack of substantive interpretation of each subdimension and great difficulty in model fitting.

In this section, we approach the task of modeling multidimensional ideal points via hierarchical model with *votesets*. Specifically, we directly group the cases into *votesets* defined by the general issue areas: civil liberties (which encompasses liberal versus conservative), economic activities (which encompasses New Deal versus Laissez Faire economic policies) and federalism (which encompasses federal versus state power). This allows us to estimate the issue-specific ideal points and the corresponding variance-covariance structure.

The vote records of the US Supreme Court Justices are extracted from the Original U.S. Supreme Court Judicial Database compiled by Harold Spaeth. This dataset includes all court cases and the vote results from 1953 to 2003. There were 29 justices and 3069 cases with non-unanimous decisions debated during this period.

In Spaeth's original database, 13 issue areas are defined based on the contents of the cases. We group them into three *votesets* of major categories: economic activities (including 543 cases related to economic activity, federal taxation, interstate relations and labor unions), civil liberties (including 1180 cases related to civil rights, criminal procedures, due process, first amendment and privacy) and federalism (including 325 cases related to attorneys, federalism and judicial power).

To understand whether the Justices have different ideological values within different issue areas, we fit five models based on different definitions of the *votesets*: considering all issue areas as one *voteset*, combining any two of the issue areas as a *voteset* and leaving the other one as another *voteset*, and treating each issue area as a *voteset*. The models and their performances are presented in Table 3.4. We can see that a 2-*voteset* model which combines issue areas economic activities and political institutions as one *voteset* and leaving cases concerning civil liberties as another *voteset* fits the data best.

Furthermore, our model reveals that there is considerable variation of the justices' ideology toward civil liberties (variance=1.17) compared to the areas of eco-

Table 3.4: Deviance Information Criterion and Mean Absolute Predictive Errors of 5 models.

votesets	DIC	$\bar{D}$	p_D	MAPE
one *voteset*	21348.4	17309.3	4039.1	0.2344
(civil and political)/economic	21123.3	17045.9	4077.4	0.2305
(economic and civil)/political	23696.2	19633.7	4062.5	0.2688
(economic and political)/ civil	20988.9	16979.8	4019.1	0.2297
economic/civil/political	22996.8	18884.6	4112.2	0.2307

nomic/political *voteset* (variance=0.36). In Figure 3.2, we plot the rank orders of the justices in the two *votesets*. Largely, the ideology of most justices remained consistent in both *votesets* and the correlation between their rank orders is 0.8. Nevertheless there are exceptions. For example, Judge Clark, an avid promoter of the New Deal economic policy, is estimated as having a moderate point of view toward civil liberties issues. On the other hand, the model shows that Justices Reed and Minton had very conservative view when deciding cases of civil liberties but were moderate justices when debating cases regarding economic activities and political institutions. The findings about these individual justices are consistent with existing anecdotal research on judicial behavior.

3.7 Discussion and Remarks

A unified approach to modeling complex legislative behaviors through *votesets*, *allysets* and *tactsets* in ideal point estimation is presented in this paper. Compared with alternative methods, this model directly speaks to a missing connection between the

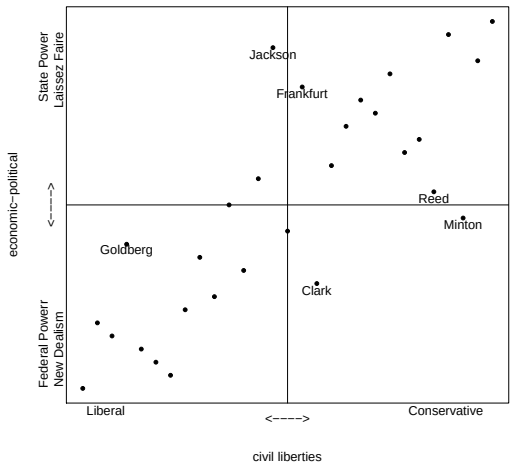

Figure 3.2: Ideal point estimates of the Supreme Court justices in different issue areas.

questions of substantive interest and statistical modeling of ideal points: modeling correlated voting behavior via dependence structures that are characterized by "covariate" and contextual variables. Hence it offers a more practical solution and is very intuitive. Guided by this framework, researchers of legislative and judicial behavior are empowered to test various propositions of formal political theories via the constructions of *votesets*, *allysets* and *tactsets* based on their substantive knowledge.

Indeed, the model we proposed in Section 3 represents a general form of modeling hierarchical structures in ideal point estimations, and can be easily generalized or reparameterized. For example, when there are independent voters, one can model the true policy position b_j^0 in absence of *allysets* effects,

$$\begin{aligned} t_{ij} &= a_j\theta_{ik} - (b_j^0 + \varphi_{lj}) + \epsilon_{ij} \\ \varphi_{lj} &\sim \mathrm{N}(0,\, \sigma_l^2) \end{aligned} \tag{3.6}$$

where φ_{lj} is the *allyset* induced effect and is equal to zero for independent voters, and

σ_ls actually quantify the exact party effects. Moreover, one can model time-varying party influence through *votesets* defined by time periods and interacting *allyset* effects and *votesets* effects (Lu & McFarland, 2007). Similar to the reparameterization (6), if we know only a subset of items violates the local item independence assumption, we can group them into *votesets* and model main ideal points of each individual, allowing occasional deviations within *votesets*.

$$\begin{aligned} t_{ij} &= a_j(\theta_i^0 + \gamma_{ik}) - b_j^0 + \epsilon_{ij} \\ \gamma_{ik} &\sim \mathrm{N}(0,\, \sigma_k^2) \end{aligned} \tag{3.7}$$

where γ_{ik} is the *voteset* induced effect and is equal to zero for independent items and σ_k actually quantifies the deviation from main ideological dimension of each *voteset*. For example, one can test whether Justice Rehnquist influences the decisions made during his tenure as Chief Justice or the Conditional Government hypothesis of conditional party influence. In reparameterization (7) the model specification is a close variation of Bradlow, Wainer and Wang's "testlet effects model" in the field of education testing (1999).

Moreover, *tactset* is a very flexible concept which allows researchers to model and test transitory collaborative voting behaviors such as strategic voting, vote trades and agenda setting.

$$t_{ij} = a_j(\theta_i + \delta_m) - b_j + \epsilon_{ij}$$

Take the example of the two *tactsets* as our illustrating example in Section 1, one is represented by the $\star$ votes and the other by the $*$ votes. We can examine whether the members of these two *tactsets* trade votes for more favorable vote outcomes in specific bills. A formal test of $\delta_1 > 0$ and $\delta_2 < 0$ is equivalent to testing the existence of a conservative shift among members of *tactset* 1 in trading for a liberal shift among

members of *tactset* 2. Not shown in this paper, simulation studies have been carried out for models with *tactsets* and consistent results are established.

Chapter 4

Variable Selection for Linear Mixed Effect Model

4.1 Introduction

Clustered data is a common phenomenon in modern data analysis. For example, in social surveys, the individual respondents are often clustered under city blocks, neighborhoods or other geographical regions. Another example can be seen in longitudinal studies where repeated measurements under the same subject are taken over time. Mixed effect models are widely used statistical tools to deal with clustered data (see for examples, Goldstein, 2002; Bryk and Raudenbush, 2001). In this paper we aim to study the problem of variable selection and parameter estimation for linear mixed effect models.

In mixed effect models, it is assumed that the unobserved heterogeneity at cluster level causes intra-cluster correlation between the responses, and hence the mean level of the responses and/or the effects of the covariates can vary across clusters. Fixed effects and random effects are used to model such intra-cluster correlation. The key

difference between fixed effect and random effect is that the former assumes unobserved heterogeneity at cluster level is constant while the latter assumes such quantity is random. Hence the estimation of the fixed effects concerns the actual sizes of the cluster-specific effects. When the number of clusters is large, the number of fixed effect coefficients increases rapidly. Conversely, for random effects researchers are more interested in the distribution rather than the actual sizes of random effect coefficients. Random effects are often assumed to follow a zero-mean multivariate normal distribution, and its covariance matrix becomes our key interest since it summarizes the intra-cluster correlation. When the number of the random effect components is large, the estimation of random effects in a mixed effect model involves a high dimensional covariance matrix that can greatly increase computational instability. Since mixed effect models are of high dimension, identifying the significant fixed effect coefficients and the effective components of random effects is very important for applied researchers to build more interpretable models and to ease the computational burden.

Traditionally, variable selection for mixed effect models has relied on p value-based stepwise deletion, or more elaborately, the Akaike's information criterion (Akaike, 1973), the FPE_λ method (Shibata, 1984), and Mallow's C_p (Mallow, 1973). However, these procedures ignore stochastic errors inherited through the process of variable selection. Hence, the estimators based on these variable selection procedures suffer from lack of stability and it is hard to understand their theoretical properties (Breiman, 1996). Alternatively, the Bayesian information criterion (Schwartz, 1978) and Generalized information criterion (Nishii, 1984; Rao and Wu, 1989) are used as consistent variable selection procedures for fixed effect parameters, but Pu and Niu (2006) found that these procedures perform poorly in selecting random effect components. Moreover, all of these variable selection procedure involve a combinatorial optimization problem which is NP-hard with computational time increasing exponentially with the number of parameters. (see comments in Fan and Li, 2004). Hence

it is not feasible to apply these procedures to the complete set of the candidate models when the number of parameters is large.

To address the weakness of traditional variable selection procedures, recent work has focused on selecting variables simultaneously with model estimation using data oriented penalty functions. For examples, the bridge regression (Frank and Friedman, 1993), the least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996, 1997), and the smoothly clipped absolute deviation penalty (SCAD) (Fan and Li, 2001). Among the alternatives, the SCAD penalty function has some oracle properties such that the estimators based on which converge to the true model while others are only shrinkage estimators. Moreover, Fan and Peng (2004) established the asymptotic properties when the number of parameters increases with sample size.

When random effects are not subject to selection, the penalty method for variable selection problem in linear mixed effect is straightforward. One can use a penalized likelihood estimation approach. When random effects are subject to selection, the problem becomes more complicated as the estimation of the covariance matrix involves a constrained optimization problem that is close to or on the boundary of the parameter space. In this situation, for most optimization procedures such as Newton-Raphson and the EM algorithm, the convergence can be slow and often fails. Only recently, Krishna (2008) developed a restricted EM algorithm that uses the adaptive LASSO (Zou, 2006) to estimate and select linear mixed model under the penalized likelihood framework.

In this paper, we aim to develop an optimization-free variable selection procedure for linear mixed effect models. To ease the burden of computation, we propose a simple iterative procedure that takes advantage of the partial consistency property of random effects. This approach has another advantage is that it allows us to select effective random effect components by penalizing random effect coefficients in groups. Antoniadis and Fan (2001) pointed out that selecting variable based on the informa-

tion of a group of variables will lead to better thresholding decision rules and faster convergence. As our simulation and theoretical results will show, this procedure selects both fixed effect and random effects consistently, and gives unbiased estimates. As sample size becomes large, the procedure has some oracle properties. Although our analysis is limited to linear mixed effect models, it provides important insights to generalized linear mixed effect models.

The rest of this paper is organized as follows: In Section Two we present a simple iterative procedure that can effectively estimate linear mixed effect model without burdensome optimization. In Section Three, we adapt this procedure to select random effect and fixed effect components simultaneously during estimation. Simulation results and an example of data analysis will be presented in Section Four. The paper ends with a discussion and future research directions.

4.2 Variable selection and estimation in linear mixed effect model

To avoid constrained optimization problem, we hereby propose to select variable and estimate parameters for linear mixed effect models based on a simple iterative procedure. We first describe how we can use this procedure to estimate linear mixed effect models and the proposed estimators can achieve satisfactory sampling properties under mild conditions. Then we extend this procedure so that it also selects the effective components of fixed effects and random effects during model estimation.

Consider the linear mixed effect model (LMM) that was originally introduced by Laird and Ware (1982). For each cluster i,

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \epsilon_i \quad i = 1, \ldots, m, \tag{4.1}$$

where $\mathbf{Y}_i$ is a vector of dependent variables of length n_i, the elements of Y_i are

assumed to be independent across clusters, but correlated within the cluster. $\mathbf{X}_i$ is a n_i by p matrix of covariates whose effects are assumed to fixed, $\boldsymbol{\beta}$ is a $p \times 1$ vector of corresponding fixed effect coefficients. To simplify the notations, we allow $\mathbf{X}$ to include both the traditional sensed covariates whose effects are constant across clusters and the cluster-specific fixed effects. $\mathbf{Z}_i$ is a n_i by q matrix of covariates whose effects are assumed to be random across clusters and $\mathbf{b}_i$ is a $q \times 1$ vector of random effect coefficients. ϵ_i is a vector of residuals of length n_i that is independent of $\mathbf{X}_i$, $\mathbf{Z}_i$ and $\mathbf{b}_i$. Moreover,

$$\begin{aligned} \epsilon_i &\sim N(0, \sigma^2 I_{n_i}), \\ \mathbf{b}_i &\sim N(0, \sigma^2 D), \\ \mathbf{Y}_i &\sim N(X_i\boldsymbol{\beta}, \sigma^2 V_i), \end{aligned}$$

where D is a $q \times q$ nonnegative definite matrix, and $V_i = I_{n_i} + Z_i^T D Z_i$. Since D characterizes the variation across groups, it is also called the variance component of the model.

4.2.1 An iterative procedure to estimate LME

Inspired by Sun, Zhang and Tong (2007), we consider the following two-step iterative procedure to estimate fixed and random effects. We start with initial values $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_0 = [\sum_i (\mathbf{X}_i^T \mathbf{X}_i)]^{-1} [\sum_i \mathbf{X}_i^T \mathbf{Y}_i]$.

Step 1: predict the residuals given $\hat{\boldsymbol{\beta}}$ for group i,

$$\mathbf{u}_i = \mathbf{Y}_i - \mathbf{X}_i \hat{\boldsymbol{\beta}},$$

for $i = 1, \ldots, m$ we can estimate

$$\hat{\mathbf{b}}_i = (\mathbf{Z}_i^T \mathbf{Z}_i)^{-1} \mathbf{Z}_i^T \mathbf{u}_i, \tag{4.2}$$

and residual $\mathbf{e}_i = \mathbf{u}_i - \mathbf{Z}_i\hat{\mathbf{b}}_i$. Based on $\mathbf{e}_i$ and $\hat{\mathbf{b}}_i$, we propose an estimator of σ^2,

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^m \mathbf{e}_i^T\mathbf{e}_i}{(n-qm)}, \tag{4.3}$$

and an estimator of D,

$$\hat{D} = \frac{\sum_{i=1}^m \hat{\mathbf{b}}_i\hat{\mathbf{b}}_i^T}{m\hat{\sigma^2}} - \frac{\sum_{i=1}^m (\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}}{m}. \tag{4.4}$$

The first term of (4.4) appears to be a naïve estimator of D. But if we look closely,

$$\hat{\mathbf{b}}_i = (\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}\mathbf{Z}_i^T\mathbf{u}_i = \mathbf{b}_i + (\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}\mathbf{Z}_i^T\epsilon_i.$$

This leads to

$$\begin{aligned}\sum_{i=1}^m \mathbf{b}_i\mathbf{b}_i^T &= \sum_{i=1}^m \hat{\mathbf{b}}_i\hat{\mathbf{b}}_i^T + \sum_{i=1}^m (\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}\mathbf{Z}_i^T\epsilon_i\epsilon_i^T\mathbf{Z}_i(\mathbf{Z}_i^T\mathbf{Z}_i)^{-1} + \sum_{i=1}^m (\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}\mathbf{Z}_i^T\epsilon_i\mathbf{b}_i^T \\ &\quad + \sum_{i=1}^m \mathbf{b}_i\epsilon_i^T\mathbf{Z}_i(\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}.\end{aligned}$$

As Sun, Zhang and Tong (2007) pointed out, the last two terms are of order $Op(m^{1/2})$, hence

$$\begin{aligned}m^{-1}\sum_{i=1}^m \mathbf{b}_i\mathbf{b}_i^T &\approx m^{-1}\left\{\sum_{i=1}^m \hat{\mathbf{b}}_i\hat{\mathbf{b}}_i^T - \sum_{i=1}^m (\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}\mathbf{Z}_i^T\epsilon_i\epsilon_i^T\mathbf{Z}_i(\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}\right\} \\ &\approx m^{-1}\left\{\sum_{i=1}^m \hat{\mathbf{b}}_i\hat{\mathbf{b}}_i^T - \sum_{i=1}^m \sigma^2(\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}\right\}.\end{aligned}$$

Substituting σ^2 by $\hat{\sigma}^2$, we obtain the estimator of D in (4.4).

Step 2: given $\hat{D}$, now we can estimate $\hat{\boldsymbol{\beta}}$ based on generalized least squares.

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{W}X)^{-1}\mathbf{X}^T\mathbf{W}\mathbf{y}, \tag{4.5}$$

where $\mathbf{W}$ is a block diagonal matrix with diagonal elements $(I_{n_i} + \mathbf{Z}_i\hat{D}\mathbf{Z}_i^T)^{-1}$, $i = 1, \ldots, m$.

To achieve numerically stable estimates of $\hat{\sigma}^2$, $\hat{\boldsymbol{\beta}}$ and $\hat{D}$, we can iterate between step 1 and step 2 until convergence.

4.2.2 Asymptotic Properties

The estimators we proposed in Section 4.2.1 have been mentioned in several papers in various contexts (for examples, Sun, Zhang and Tong (2007) and Demindeko (2006).) In this section, we systematically show that the estimators of β, D are $\sqrt{n}-$consistent. To make the presentation clearer, we introduce the following notations,

$$c_1 = \lim_{m\to\infty} \frac{n}{n-qm} \tag{4.6}$$

$$c_2 = \lim_{m\to\infty} \frac{n}{m} \tag{4.7}$$

$$\Gamma = \lim_{m\to\infty} \frac{1}{m}\sum_{i=1}^{m} \mathrm{E}[(\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}], \tag{4.8}$$

$$\Delta_2 = \lim_{m\to\infty} \frac{1}{m}\sum_{i=1}^{m} \mathrm{E}[\{(\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}\}\otimes\{(\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}\}] \tag{4.9}$$

$$\Delta_3 = \lim_{m\to\infty} \frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{n_i} \mathrm{E}[\mathrm{vec}\{(\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}Z_{ij}Z_{ij}^T(\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}\}\mathrm{vec}^T\{(\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}Z_{ij}Z_{ij}^T(\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}\}]$$

$$\Delta_4 = \mathrm{vec}\left\{\mathbf{D}+\frac{1}{m}\sum_{i=1}^{m}(\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}\right\}\mathrm{vec}\left\{\mathbf{D}+\frac{1}{m}\sum_{i=1}^{m}(\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}\right\}^T \tag{4.10}$$

$$\gamma = \lim_{m\to\infty}(n-qm)^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n_j}\mathrm{E}[Z_{ij}^T(\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}Z_{ij}]^2 - c_1q/c_2+1 \tag{4.11}$$

and

$$\Delta_1 = \begin{pmatrix} \mathbf{D}\otimes\Gamma_{(1)}+\Gamma\otimes\mathbf{D}_{(1)} \\ \vdots \\ \mathbf{D}\otimes\Gamma_{(q)}+\Gamma\otimes\mathbf{D}_{(q)} \end{pmatrix}$$

where $\Gamma_{(r)}, \mathbf{D}_{(r)}(r=1,\ldots,_q)$ denote the rth row of $\mathbf{D},\Gamma$, respectively, $\otimes$ is the kronecker product, vech($\mathbf{A}$) denotes the vector consisting of all elements on and below the diagonal of the matrix, and vec($\mathbf{M}$) denotes the vector by simply stacking the column vectors of the matrix $\mathbf{M}$ below one another. Obviously there exists a unique $q^2\times q(q+1)/2$ matrix R_q such that $\mathrm{vec}(\mathbf{A}) = R_q\mathrm{vech}(\mathbf{A})$.

Under some mild conditions, we have the following results.

Lemma 4.2.1 *Under the regularity conditions (A)-(D),*

$$n^{1/2}\left\{\hat{\sigma}^2-\sigma^2\right\} \xrightarrow{D} \mathcal{N}(0, 2\sigma^4(1+\gamma)c_1 + Var(\epsilon_{11})\gamma c_1).$$

Proposition 4.2.1 *Under the regularity conditions (A)-(D), given a $\sqrt{n}-$consistent estimator $\hat{\mathbf{D}}$ of $\mathbf{D}$, for the generalized least square estimator of $\boldsymbol{\beta}$,*

$$\hat{\boldsymbol{\beta}} = argmin \sum_{i=1}^{m} (\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta})^T (\mathbf{I} + \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i')^{-1}(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta}),$$

we have

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} \mathcal{N}(0, \Sigma_\beta),$$

where

$$\Sigma_\beta = \lim_{m\to\infty} \sigma^2 \left(\sum_{i=1}^{m} \mathbf{X}_i'(\mathbf{I} + \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i')^{-1}\mathbf{X}_i \right)^{-1}.$$

Proposition 4.2.2 *Under the regularity conditions (A)-(D), given a $\sqrt{n}-$consistent estimate of β, for the estimate of $\mathbf{D}$ by (4.4), we have*

$$\sqrt{n}\left\{ vec(\hat{\mathbf{D}} - \mathbf{D}) \right\} \xrightarrow{D} \mathcal{N}(0, (R_q^T R_q)^{-1} R_q^T \Delta R_q (R_q^T R_q)^{-1} c_2),$$

where

$$\begin{aligned} \Delta &= E\{\mathbf{b}_1\mathbf{b}_1^T \otimes \mathbf{b}_1\mathbf{b}_1^T\} - vec(\Sigma)vec^T(\Sigma) + \sigma^2\{\Sigma\otimes\Gamma + \Gamma\otimes\Sigma + \Delta_1\} \\ &\quad +2\sigma^4\{\Delta_2 - \Delta_3 + c_1/c_2(1+\gamma)\Delta_4\} + var(\varepsilon_{11}^2)\{\Delta_3 + c_1/c_2\gamma\Delta_4\}. \end{aligned}$$

4.3 Selecting Effective Fixed and Random Effects components

Since the procedure we proposed in Section 4.2.1 is an optimization-free one, it enjoys great computational stability even when the covariance matrix is near singular. In

this section, we consider selecting the effective components of fixed and random effects in linear mixed effect model via the penalty function SCAD.

Now suppose in model (4.1), some components of $\boldsymbol{\beta}$ are zero, and some random effects are zero such that the corresponding diagonal elements of D are zero. Without loss of generality, we write

$$\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{10}^T, \boldsymbol{\beta}_{20}^T)^T$$

where $\boldsymbol{\beta}_{20} = \mathbf{0}$, and $\text{diag}(D_0) = (\mathbf{d}_{10}^T, \mathbf{d}_{20}^T)^T$, where $\mathbf{d}_{20} = \mathbf{0}$, and corresponding rows and columns of D_0 are zero as well.

To be able to simultaneously select the nonzero components of fixed effects and random effects during the estimation, we adjust the above two-step estimating procedures such that the small fixed effect coefficients will be shrunk to zero and the effective dimension of D will be correctly identified.

4.3.1 An Iterative Procedure to Select and Estimate LME

Step 1: First observe that if the kth random effects component is effectively absent, then the (k, k) diagonal element of D is zero, and so are the elements of the corresponding kth row and kth column since the correlation between the kth random effect and other components of random effects is zero. Hence we expect the estimate $\hat{D}$ as given in (4.4) will be close to zero as well. Using this fact, we consider shrinking the corresponding random effect coefficients $\mathbf{b}_{ik}$, $i = 1, \ldots, m$ to zero if their variance is estimated to be sufficiently close to zero.

Hence we propose to estimate $\mathbf{b}_i$, for each i, $i = 1, \ldots, m$, by minimizing the following penalized least squares

$$\frac{1}{2}(\mathbf{u}_i - \mathbf{b}_i\mathbf{Z}_i)^T(\mathbf{u}_i - \mathbf{b}_i\mathbf{Z}_i) + \sum_{k=1}^{q} np_\xi(c_k), \tag{4.12}$$

where $c_k = \sqrt{D_{kk}}$, D_{kk} is the kth diagonal element of the covariance matrix D. $p_\xi(\theta)$ is the smoothly clipped absolute deviation penalty function (SCAD) by Fan and Li (2001), where

$$p'_\xi(\theta) = \xi\Big\{I(\theta \le \xi) + \frac{(a\xi - \theta)_+}{(a-1)\xi} I(\theta > \xi)\Big\}, \tag{4.13}$$

for some $a > 2$ and $\theta > 0$. ξ is the tuning parameter for penalty function $p(\cdot)$. As Fan and Li pointed out, the SCAD function is singular at the origin and does not have a continuous secondary derivative. To solve this penalized least squares, one can locally approximate the penalty function by its quadratic function when $c_k \neq 0$,

$$[p_\xi(c_k)]' \approx \{p'_\xi(c_{k0})/c_{k0}\}c_k, \quad \text{for } c_{k0} \approx c_k.$$

In other words,

$$p_\xi(c_k) \approx p_\xi(c_{k0}) + \frac{1}{2}\{p'_\xi(c_{k0})/c_{k0}\}(c_k^2 - c_{k0}^2).$$

Consequently, the solution to (4.12) can be updated based on the following ridge regression

$$\mathbf{b}_i^* = (\mathbf{Z}_i^T\mathbf{Z}_i + n\Sigma_\xi(\mathbf{c_0}))^{-1}\mathbf{Z}_i^T\mathbf{u}_i, \quad i = 1, \ldots, m. \tag{4.14}$$

where $\mathbf{c_0} = \text{diag}(\mathbf{c_1}, \ldots, \mathbf{c_q})$ and $\Sigma_\xi = \text{diag}(\frac{p'_\xi(c_{10})b_{i10}\text{sgn}(b_{i10})}{mc_{10}}, \ldots, \frac{p'_\xi(c_{q0})b_{iq0}\text{sgn}(b_{iq0})}{mc_{q0}})$.

An estimator of D is then,

$$D^* = \frac{\sum_{i=1}^m \mathbf{b}_i^*\mathbf{b}_i^{*T}}{m\hat{\sigma}^2} - \frac{\sum_{i=1}^m (\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}}{m}. \tag{4.15}$$

If the variance of the kth random effect c_j is estimated to be small, then we expect the solution $\hat{b}_{ik}$ in (4.14) will shrink to zero. This is true for all $i = 1, \ldots, m$. Correspondingly, the corresponding diagonal elements and the corresponding rows and columns of D^* in (4.15) will be estimated to be zero.

Comments: Our approach of shrinking a group of random effect coefficients together is closely related to the blocked-wise penalized functions that were discussed in Antoniadis and Fan (2001). In particular, the blocked-wise penalized least squares problem takes the following form,

$$\|Z - \theta\|^2 + p_\lambda(\|\theta\|)$$

. They argue that whenever applicable, to shrink the coefficients in groups will make the thresholding decision more accurate and improve the convergence rate since the information within a group is bigger. This idea is also seen in more recent work such as group LASSO (Yuan and Lin, 2006).

Step 2: The selection of fixed effects given random effect variance estimates $\hat{D}$ is simply the solution to the penalized weighted least squares

$$\frac{1}{2}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T\mathbf{W}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + n\sum_{k=1}^{p} p_\lambda(|\beta_k|). \tag{4.16}$$

Similarly, we can use a local quadratic approximation of $p_\lambda(|\beta_k|)$ and update $\boldsymbol{\beta}^*$ based on the following ridge regression,

$$\boldsymbol{\beta}^* = (\mathbf{X}^T\mathbf{W}X + n\Sigma_\lambda(|\boldsymbol{\beta}_0|))^{-1}\mathbf{X}^T\mathbf{W}\mathbf{y}, \tag{4.17}$$

where λ is the tuning parameter for the penalty function, and

$$\Sigma_\lambda(\boldsymbol{\beta}_0) \;=\; \text{diag}\left[p'_\lambda(|\beta_{10}|)/|\beta_{10}|, \ldots, p'_\lambda(|\beta_{p0}|)/|\beta_{p0}|\right].$$

To achieve numerically stable estimates of $\boldsymbol{\beta}^*$ and D^*, we can iterate between step 1 and step 2 until convergence.

4.3.2 Asymptotic Properties

We can show that the estimators of D and β given in (4.15) and (4.17) are consistent and have some oracle properties.

Theorem 4.3.1 *Under the regularity conditions (A)-(D), given a $\sqrt{n}-$consistent estimate D^* of $\mathbf{D}$, if $\sqrt{n}\lambda_n \to \infty$ and $\lambda_n \to 0$ as $n \to \infty$, then there is a local minimizer $\boldsymbol{\beta}^*$ of (4.16) such that*

$$\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\| = O_p(1/\sqrt{n}),$$

and this minimizer must satisfy

(a) Sparsity: $\boldsymbol{\beta}_2^* = 0$.

(b) Asymptotic normality:

$$\sqrt{n}(\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_{01}) \xrightarrow{D} \mathcal{N}(0, \Sigma_{\beta_{01}})$$

where

$$\Sigma_{\beta_{01}} = \lim_{m\to\infty} \sigma^2 \left(\sum_{i=1}^{m} \mathbf{X}_{i1}'(\mathbf{I} + \mathbf{Z}_i \mathbf{D}^* \mathbf{Z}_i')^{-1} \mathbf{X}_{i1} \right)^{-1}.$$

Theorem 4.3.2 *Under the regularity conditions (A)-(D), given a $\sqrt{n}-$consistent estimate of β, if $\sqrt{n/log(n)}\xi_n \to O(1)$ as $n \to \infty$, then then there is a local minimizer $\mathbf{b}_i^*, i = 1, \ldots, m$ of (4.14) such that for the estimate of $\mathbf{D}^*$ by (4.15) we have*

(a) Sparsity: $\mathbf{d}_2^* = 0$.

(b) Asymptotic Normality:

$$\sqrt{n}\left\{vech(\mathbf{D}_1^* - \mathbf{D}_{01})\right\} \xrightarrow{D} \mathcal{N}(0, (R_q^T R_q)^{-1} R_q^T \Delta R_q (R_q^T R_q)^{-1} c_2)$$

where $\mathbf{Z}_i, \mathbf{b}_i, q$ are replaced by $\mathbf{Z}_{i1}, \mathbf{b}_{i1}$ and q_1 in (4.6)—(4.11) and the definition of R_q.

4.3.3 Tuning Parameter Selection and Thresholding

To implement our variable selection procedure in Section 4.2, we need to consider the choice of tuning parameter λ and ξ. Theoretically, we need $\lambda \to 0$ and $\xi \to 0$

as $n \to 0$ but faster than $O(n^{-1/2})$ in order to consistently select fixed and random effects. In practice, the tuning parameter can be selected based on data oriented method. Following Fan and Li (2001), Wang, Li and Tsai (2007), we consider the following three criteria,

1. generalized cross-validation criterion

$$\text{argmin}_\lambda \text{GCV}_\lambda = \frac{\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2_{\mathbf{W}}}{n(1 - \text{Df}_\lambda/n)},$$

2. the AIC criterion

$$\text{argmin}_\lambda \text{AIC}_\lambda = \log\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2_{\mathbf{W}} + 2\text{Df}_\lambda/n,$$

3. the BIC criterion

$$\text{argmin}_\lambda \text{BIC}_\lambda = \log\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2_{\mathbf{W}} + \text{Df}_\lambda \log(n)/n.$$

where $\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2_{\mathbf{W}}$ is the model error for linear mixed model, $\mathbf{W}$ is a block diagonal matrix with diagonal elements $(I_{n_i} + \mathbf{Z}_i \hat{D} \mathbf{Z}_i^T)^{-1}$,, $i = 1, \ldots, m$. Wang, Li and Tsai (2007) argued that the BIC criterion is an optimal and consistent procedure to select the tuning parameter for linear regression, while GCV and AIC criteria tend to overfit the model. We expect this argument holds for our application as well.

The degree of freedom is hard to determine in linear mixed effect model. Here we adopt Hodges & Sargent (2001)'s formula to calculate the degree of freedom. We write model (4.1) by adding a block of "pseudo data"

$$\mathbf{Y}^* = \mathbf{U}\delta + \mathbf{e},$$

where

$$\mathbf{Y}^* = \begin{pmatrix} \mathbf{Y} \\ 0_{qm} \end{pmatrix}, \quad \delta = \begin{pmatrix} \boldsymbol{\beta} \\ \mathbf{b} \end{pmatrix}, \quad \mathbf{U} = \begin{pmatrix} \mathbf{X} & \mathbf{Z} \\ 0 & -\Delta \end{pmatrix}, \mathbf{e} = \begin{pmatrix} \epsilon \\ \mathbf{b} \end{pmatrix},$$

where $\Delta^T\Delta = G^{-1}$, and G is a $qm \times qm$ block diagonal matrix with diagonal element D. Then we can obtain a quasi "Hat" matrix H_1 for linear mixed model,

$$H_1 = \begin{pmatrix} \mathbf{X} & \mathbf{Z} \end{pmatrix} (\mathbf{U}^T\mathbf{U})^{-1} \begin{pmatrix} \mathbf{X}^T \\ \mathbf{Z}^T \end{pmatrix}.$$

The effective degree of freedom is then trace(H_1).

4.4 Simulation Studies and Real Data Analysis

In this section, we conduct a set of simulation studies to assess the performance of the proposed variable selection and estimation procedure for linear mixed effect model. A real data analysis will also be conducted. We are particularly interested in model performance in the following aspects: whether the correct subsets of fixed effects and random effects can be correctly selected; whether the parameter estimates are unbiased and efficient in small to medium sample sizes; and when the true models are ascertained, whether the iterative method has comparable sample properties to maximum likelihood method.

4.4.1 Simulation I

In the first set of simulations, we adopt the examples in Krishna (2008). There are two scenarios:

1. Example 1: Consider $m = 30$ subjects, $n_i = 5$ observations per subject. There are 9 fixed effects to be considered, the true value of coefficients are $\boldsymbol{\beta} = [0, 1, 1, 0, 0, 0, 0, 0, 0]$. For random effects, we consider 4 dimensions, with

the true covariance matrix

$$D = \begin{pmatrix} 9 & 4.8 & 0.6 & 0 \\ 4.8 & 4 & 1 & 0 \\ 0.6 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

The model variance $\sigma^2 = 1$. Furthermore, covariates $\mathbf{X}$ are generated from a uniform $(-2, 2)$ distribution, along with a vector of $\mathbf{1}$'s for the subject-specific intercept. The values of $\mathbf{Z}$ are taken to be the values of the first four columns of $\mathbf{X}$.

2. Example 2: The set up of the second example is the same as the first, except the number of the subjects increases to $m = 60$ and the number of the observations per subject increases to $n_i = 10$.

For each example, we randomly draw 200 samples and apply the proposed variable selection and estimation procedure to these data sets. In Table 4.1, we summarize the performance of the proposed iterative procedure under different tuning parameters. We notice that as sample size grows, the procedure selects the correct fixed and random effect components with increasing accuracy. Due to the benefit of group selection, the selected random effect components in particular quickly converge to the true model. For different tuning parameter choices, we can see that in general the BIC criterion outperforms the other tuning parameter choices. Both the false positive rate and the false negative rate are smaller for the models selected based on the BIC criterion. The average model size is closer to that of the true model as well.

We also compare the the percentage of the models that are correctly identified by our procedure in comparison with Krishna's Table (3.1). In Table 4.2, We can see as the sample size increases, the performance of our method improves dramatically. With a fixed size of 600, random effect selection has nearly 100% accuracy and the fixed

Tuning	FPR%	FNR%	Model Size	FPR%	FNR%	Model Size
Fixed Effects	Example 1			Example 2		
BIC	21.5	9.9	2.26	1.5	1.9	2.10
AIC	17	11.0	2.43	1.5	3.3	2.20
GCV	20.5	10.1	2.30	1.5	3	2.18
$\sqrt{\log(n)/n}$	21	15.6	2.67	1.5	4.1	2.26
Random Effects						
BIC	27	6	2.25	0	0	3
AIC	25	12	2.37	0	0	3
GCV	26	6	2.28	0	0	3
$\sqrt{\log(n)/n}$	33	7	2.09	0	0	3

Table 4.1: Performance of fixed and random effect selection. "FPR%" is the average false positive rate which is defined as the percentage of the coefficients that are incorrectly estimated to be nonzero. "FNR%" is the average false negative rate that is the percentage of the coefficients that are incorrectly estimated to be zero. "Model size" reports the average size of nonzero fixed effect coefficients and nonzero random effect components.

effect selection (using GIC as tuning parameter criterion) outperforms all the other existing approaches. As a matter of fact, the overall model selection performance is partly impacted by the simulation set up. The first random effect is assumed a large variance of 9 which causes large uncertainty in estimating the random effect coefficients $\mathbf{b}_1$ and in turn affects the estimation and selection of the first fixed effect coefficient in our iterative procedure. A detailed look of the fixed effect selection reveals that our method successfully selects all but the first fixed effect in the 200 simulations we conducted.

On the other hand, our model performs relatively unsatisfactorily with low sample size, especially when the number of observations per group is low compared to the number of the random effect coefficients need to be estimated. Since our method is conditional on the estimated individual random effect coefficients, when the number of within group observation is low compared to the number of random effect coefficients that need to be estimated, it is understandable that it does not perform as well as the penalized likelihood approaches where only the marginal distribution of the random effects is involved.

Method	Tuning	%Correct	%CF	%CR	%Correct	%CF	%CR
		Example 1			Example 2		
Iterative method	BIC	19	49	35	86	86	100
Iterative method	AIC	21	46	35	77	77	100
Iterative method	GCV	20	49	37	79	79	100
Iterative method	$\sqrt{\log(n)/n}$	16	33	27	72	72	100
M-ALASSO	BIC	71	73	79	83	83	89
EGIC	BIC	47	56	52	48	59	53
RIC	AIC	19	21	62	31	34	74
RIC	BIC	59	59	68	77	79	81
Stepwise	AIC	17	21	62	26	28	74
Stepwise	BIC	51	53	68	68	69	81
ALASSO	AIC	21	24	62	39	41	74
ALASSO	BIC	62	63	68	74	75	81

Table 4.2: Comparing the model selection performance of the proposed iterative method with other existing methods. "% Correct" reports the percentage of times the correct true model was selected, "% CF" and "%RF" report the percentage of the times correct fixed effect components and random effect components are selected. The results for M-ALASSO, EGIC, RIC, Stepwise deletion and ALASSO are borrowed from Krishan (2008).

4.4.2 Simulation II

In the second simulation study, we focus on the performance of parameter estimates of our proposed methods. We consider the following 4 different scenarios: the number of clusters is either 10 or 20, and the number of observations within each cluster is either 10 or 20. For each scenario, we assume the same model structure as follows: the dimension of fixed effects is $p = 5$ with true value $\boldsymbol{\beta} = [1, 0, 1.5, 1, 0]$. The dimension of the random effects is $q = 4$ with the covariance matrix of the random effect coefficients D specified as follows,

$$D = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0.354 \\ 0 & 0 & 0 & 0 \\ 0 & 0.354 & 0 & 1 \end{pmatrix}$$

so that only the second and the fourth random effect components are significant. Furthermore, the correlation between the second and the fourth random effects is 0.5. The model variance σ^2 is assumed to be 1. Without loss of generality, the components of $\mathbf{X}$ are generated from standard normal distributions, and $\mathbf{Z}$ assumes the same values as $\mathbf{X}_1, \ldots, \mathbf{X}_4$.

For each scenario, we simulate 100 data sets and run the iterative variable selection and estimation procedure for each data set. For the purpose of comparison, we estimate three different models. First we apply the proposed iterative penalized method to select and estimate both fixed effects and random effects simultaneously. Then we estimate the model using the iterative procedure and the maximum likelihood estimation assuming the true model is known.

The number of correctly and incorrectly selected fixed effects and random effects among the 100 simulated data sets are reported in Table 4.3. We can clearly see that as both the number of clusters and the number of within cluster observations increase,

the fixed effects and random effects are selected with increasingly high accuracy.

sample size		Fixed Effects					Random Effects			
m	n_i	β_1	β_2	β_3	β_4	β_5	D_1	D_2	D_3	D_4
10	10	0	92	1	11	94	100	39	100	5
20	10	1	98	1	0	98	100	8	100	0
10	20	1	96	0	1	95	100	9	100	1
20	20	1	100	0	0	99	100	0	100	0

Table 4.3: Numbers of fixed effects and random effects that are selected to be zero in 100 simulated data sets. For β_2, β_5, D_1 and D_3, the table reports the number of parameters that are correctly selected to be zero. For β_1, β_3, β_4, D_2 and D_4, it reports the number that are incorrectly selected to be zero.

Next we examine the performance of parameter estimation of our proposed models. For each simulation set up, we present bias and median absolute deviation of the nonzero fixed effect and random effect parameters in Table 4.4. These summary statistics demonstrate that the parameter estimators based on our proposed iterative procedure possess satisfying sampling properties. For both fixed effects and random effects, the estimators are unbiased and behave as if the true model is known when sample size is large. Moreover, we can see that when the true model is known, our proposed iterative procedure performs equally well as the maximum likelihood estimation.

m	n_i	parameter	Bias			MAD		
			iter	iterO	MLEO	iter	iterO	MLEO
10	10	β_1	0.006	0.012	0.011	0.073	0.065	0.067
		β_3	-0.010	-0.003	0.000	0.064	0.065	0.064
		β_4	-0.112	0.022	0.019	0.292	0.260	0.251
		D_{22}	-0.124	-0.008	-0.002	0.275	0.135	0.130
		D_{44}	0.150	0.038	-0.040	0.437	0.321	0.313
		D_{24}	-0.120	-0.019	-0.023	0.115	0.185	0.189
20	10	β_1	-0.001	0.008	0.009	0.046	0.044	0.046
		β_3	-0.006	0.002	0.002	0.046	0.046	0.046
		β_4	-0.007	-0.010	-0.008	0.141	0.140	0.130
		D_{22}	-0.010	0.013	0.026	0.128	0.125	0.155
		D_{44}	-0.004	-0.031	-0.029	0.238	0.231	0.257
		D_{24}	-0.024	-0.005	0.008	0.117	0.109	0.112
10	20	β_1	-0.005	0.004	0.004	0.047	0.046	0.046
		β_3	-0.006	-0.008	-0.008	0.053	0.051	0.047
		β_4	0.052	0.063	0.063	0.207	0.186	0.189
		D_{22}	-0.062	-0.033	-0.033	0.183	0.160	0.141
		D_{44}	-0.006	-0.023	-0.090	0.253	0.269	0.241
		D_{24}	-0.033	-0.025	-0.027	0.188	0.185	0.164
20	20	β_1	-0.010	0.001	0.001	0.034	0.033	0.033
		β_3	0.007	0.004	0.004	0.037	0.037	0.036
		β_4	0.003	0.001	0.002	0.126	0.121	0.120
		D_{22}	0.013	0.013	0.020	0.113	0.112	0.111
		D_{44}	-0.021	-0.021	-0.040	0.201	0.201	0.210
		D_{24}	-0.003	-0.002	0.005	0.119	0.118	0.110

Table 4.4: Bias and median absolute deviation (MAD) of the significant fixed effect and random effect parameter estimates. "iter" refers to iterative variable selection and estimation method, "iterO" refers to iterative estimation method under the true model, and "MLEO" refers to MLE under the true model.

4.4.3 Real Data Analysis

In this section, we apply the proposed method to the 2004 American National Election Study. The ANES is a series of surveys that capture voters's opinions before and after each election since 1948. The outcome variable we are interested in is the feeling thermometer reading for George W. Bush. Feeling thermometer is a widely accepted way of quantifying individuals' feeling toward public figures. It mimics a physical thermometer and ranges from 0 to 100 degrees. The higher temperature an individual assigns indicates he/she feels more positive toward Bush, and vice versa.

The 2004 ANES is a national representative sample of 1212 respondents from 29 states in the US. In this analysis, we will examine what factors affect individuals' feeling toward Bush. Since such effects tend to be mediated by social and cultural contexts at the state level, we will further examine whether these effects vary across states. Hence we fit a linear mixed effect model with individuals nested under states. After removing missing data and states with too few observations, the effective sample size consists of 1156 individuals from 24 states.

Figure 4.1 shows the histogram of Bush feeling thermometer readings. We can see that there is considerable amount of variation and the distribution appears to be bimodal rather than normally distributed. Like many other social and behavioral studies, there exists a large amount of individual level heterogeneity that can not be easily captured via systematic modeling. Our results reveal that the model variance σ^2 is rather large compared to the amount of variance systematically explained by fixed effects and random effects (the intracluster correlation is only 18%). Moreover, since there could be a wide arrays of factors influencing individuals' preference toward political figures, we start with a linear mixed effect model with a large number of fixed and random effects (see Table 4.5).

In this data analysis, large model variance and a number of potentially nuisance

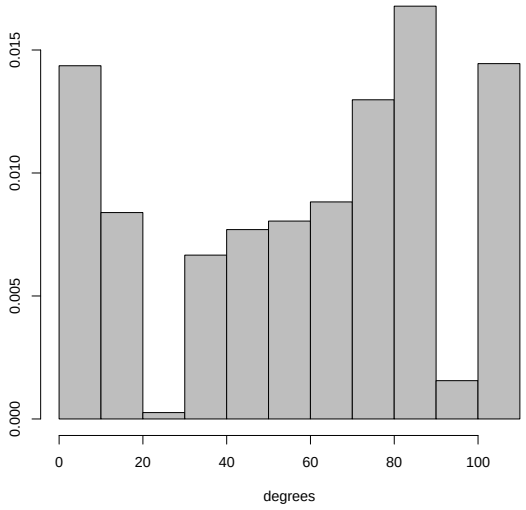

Figure 4.1: Histogram of the outcome variable "Bush feeling thermometer readings"

random effect components can pose great challenge to maximum likelihood based approaches in estimating and selecting the correct submodel. In our attempts to fit the model using commercial software such as the `xtmixed` package in STATA and the `nlme` package in R, the initial full model and its many submodels fail to converge. To tackle this problem, we apply the proposed procedure in Section 4.3 to estimate the model while shrinking the insignificant fixed and random effects to zero simultaneously. We use general cross validation method to determine the values of the tuning parameters for selecting fixed effects and random effects via the SCAD function. The results are presented in Table 4.5. Among the 14 random effects listed in Table 4.4, only two are deemed to be effective random effect components. For comparison purposes, we also fit the same model using R package `nlme` that is based on RMLE. We can see that in general our model yields estimates that are close to the R package. However, as mentioned before, since the outcome variable Bush

Fixed effects	Random Effects
intercept, age, gender, education, income, Christian, black, other, gun control, liberal view, moderate view, defense issues, abortion right, death penalty, environment issues, social trust, church attendance, health insurance, Democrat, Independent, Iraq war	intercept, gender,income, Christian, gun control, liberal view, moderate view, defense issues, abortion right, death penalty, health insurance, Democrat, Independent, Iraq war

Table 4.5: The complete lists of the candidate fixed effect and random effect components.

feeling thermometer appears to be bimodally distributed, our approach is expected to provide more robust results than the MLEs that are based on the assumption of normality.

Method	Iterative Method			`nlme` package		
Fixed effects ($\boldsymbol{\beta}$)	coefficient	s.e.	p value	coefficient	s.e.	p value
(Intercept)	48.57	3.46	0.00	46.62	3.34	0.00
age	0.06	0.04	0.12	0.09	0.04	0.02
education	-4.96	1.36	0.00	-5.23	1.33	0.00
Christian	7.32	1.65	0.00	7.62	1.76	0.00
black	-3.97	2.17	0.06	-2.52	2.01	0.21
other	3.6	2.03	0.07	4.59	1.97	0.02
liberal	-11.97	1.77	0.00	-11.68	1.75	0.00
defense	2.39	1.35	0.07	1.95	1.31	0.14
death penalty	4.17	1.44	0.00	4.62	1.43	0.00
Democrat	-24.48	2.08	0.00	-25.05	2.06	0.00
Independent	-14.21	1.73	0.00	-14.53	1.71	0.00
Iraq war	30.47	1.61	0.00	30.62	1.58	0.00
Random effects (D)						
gender	52.32			55.38		
Christian	25.91			15.44		
Covariance	-19.65			-28.27		
Model variance (σ^2)						
	438.02			442.26		

Table 4.6: Parameter estimation of the fixed effect coefficients and the random effect covariance. The first three columns report the coefficients, standard errors and p−values estimated based on the iterative variable selection and estimation procedure. The last three columns report the corresponding estimates based on R package `nlme` under the model that is selected via iterative procedure.

4.5 Discussion

In this paper, we present a simple iterative penalized procedure that selects and estimates fixed effects and random effects simultaneously. The theoretical and simulation investigations of the proposed procedure have shown that it selects the correct submodel effectively and has some oracle properties. Although in mixed effect model it is well known that the random effect coefficients can not be consistently estimated, we have demonstrated that the covariance of the random effect coefficients can be consistently estimated. Moreover, we can take advantage of this partial consistency property to select the effective component of random effects by penalizing the random effect coefficients in group. If the corresponding variance term is sufficiently small, then we shrink the entire group of random effect coefficients to zero via penalized least squares.

Our method is based on the estimation of the random effect coefficients. The cost of relying on the estimation of the random effect coefficients is that we need sufficient number of observation within each cluster. When the cluster size is small relative to the dimension of random effects, our method does not perform as well as the likelihood based approaches that only concern the marginal distribution of the data. However, in survey data analysis, the size of the clusters is typically large, so we expect this method offers a practical solution to many real data analysis problems.

In general, this method enjoys many advantages over the classical likelihood based approaches. Compared to the classical likelihood approach, this procedure has greater computational stability since it avoids the complicated constrained optimization problem of estimating a high dimensional covariance matrix that is located at the boundary of the parameter space due to the inclusion of non-existing random effects.

Moreover, since our method does not rely on multivariate normal distribution of the data, it is expected to be robust under model misspecification. In particular,

we can further relax step 2 of the iterative procedure: instead of using (penalized) weighted least squared that takes the normal covariance structure of the error terms into account, we can simply use (penalized) ordinary least squares to calculate $\boldsymbol{\beta}$ based on $\mathbf{Y}_{0i} = \mathbf{Y}_i - \mathbf{Z}_i\mathbf{b}_i$.

$$\boldsymbol{\beta} = (\mathbf{X}^T\mathbf{X} + n\Sigma_\lambda(|\boldsymbol{\beta}_0|))^{-1}\mathbf{X}^T\mathbf{Y}_0 \tag{4.18}$$

Based on simulation evidence (not shown in this paper), this distribution-free version of the iterative procedure can also select fixed effects and random effects satisfactorily. Although this procedure is less efficient when the errors are known to be normally distributed, it is more robust if the model is misspecified as it does not depend on particular information of the error structure.

Lastly, this method can be easily adapted to estimate multiple levels of hierarchical structure. To select and estimate fixed and random effects at multiple levels, we can simply condition on the random effect coefficients at lower level and partial consistency property will ensure the validity of this approach.

Appendix: Proofs

In this section, we outline the detailed proofs of the asymptotic results in previous sections.

First, we state the following regularity conditions under which the proofs are derived.

(A) $\mathrm{E}(\varepsilon_{11}^4) < \infty, \mathrm{E}\|\mathbf{b}_1\|^4 < \infty, \mathrm{E}x_i^{2s} < \infty$ and $\mathrm{E}z_j^{2s} < \infty$ where $\|\mathbf{b}_1\| = (\mathbf{b}_1^T\mathbf{b}_1)^{1/2}$, x_i denotes the ith element of $\mathbf{X}$ and z_i denotes the jth element of $\mathbf{Z}$ for $s > 2$, $i = 1, \ldots, p$, $j = 1, \ldots, q$.

(B) The elements of $\mathbf{Z}_i, i = 1, \ldots, m$ are uniformly bounded by a constant.

(C) The minimum eigenvalue of $\mathbf{Z}_i^T\mathbf{Z}_i$ and $\mathbf{X}_i^T\mathbf{X}_i, i = 1, \ldots, m$ are uniformly larger than a constant.

(D) The size of each cluster, $n_i, i = 1, \ldots, m$ is bounded by a constant, so $m = nO(1)$.

Proof of Lemma 4.2.1: It can be deduced directly from Theorem 2 of Sun, Zhang, and Tong (2007).

Proof of Proposition 4.2.1: It is not difficult to show that

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{m} \mathbf{X}_i^T(\mathbf{I} + \mathbf{Z}_i\hat{\mathbf{D}}\mathbf{Z}_i')^{-1}\mathbf{X}_i\right)^{-1} \sum_{i=1}^{m} \mathbf{X}_i^T(\mathbf{I} + \mathbf{Z}_i\hat{\mathbf{D}}\mathbf{Z}_i')^{-1}\mathbf{Y}_i.$$

Next define

$$\hat{\boldsymbol{\beta}}^* = \left(\sum_{i=1}^{m} \mathbf{X}_i^T(\mathbf{I} + \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i')^{-1}\mathbf{X}_i\right)^{-1} \sum_{i=1}^{m} \mathbf{X}_i^T(\mathbf{I} + \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i')^{-1}\mathbf{Y}_i.$$

Based on linear model theory,

$$\sqrt{n}(\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta}) \xrightarrow{D} \mathcal{N}(0, \Sigma_\beta)$$

Hence to prove Proposition 2.1, by Slutsky's lemma, we only need to prove that

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^*) = o_p(1). \tag{4.19}$$

To show (4.19), we rewrite $\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^*$ as

$$\begin{aligned}
\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^* &= \left(\sum_{i=1}^{m} \mathbf{X}_i^T(\mathbf{I} + \mathbf{Z}_i\hat{\mathbf{D}}\mathbf{Z}_i')^{-1}\mathbf{X}_i\right)^{-1} \sum_{i=1}^{m} \mathbf{X}_i^T(\mathbf{I} + \mathbf{Z}_i\hat{\mathbf{D}}\mathbf{Z}_i')^{-1}\mathbf{Y}_i \\
&\quad - \left(\sum_{i=1}^{m} \mathbf{X}_i^T(\mathbf{I} + \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i')^{-1}\mathbf{X}_i\right)^{-1} \sum_{i=1}^{m} \mathbf{X}_i^T(\mathbf{I} + \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i')^{-1}\mathbf{Y}_i \\
&= I_1 + I_2.
\end{aligned}$$

where

$$I_1 = \left(\sum_{i=1}^{m} \mathbf{X}_i^T(\mathbf{I}+\mathbf{Z}_i\hat{\mathbf{D}}\mathbf{Z}_i')^{-1}\mathbf{X}_i\right)^{-1}\left\{\sum_{i=1}^{m} \mathbf{X}_i^T(\mathbf{I}+\mathbf{Z}_i\hat{\mathbf{D}}\mathbf{Z}_i')^{-1}(\mathbf{Z}_i\mathbf{b}_i+\epsilon_i) - \sum_{i=1}^{m} \mathbf{X}_i^T(\mathbf{I}+\mathbf{Z}_i\mathbf{D}\mathbf{Z}_i')^{-1}(\mathbf{Z}_i\mathbf{b}_i+\epsilon_i)\right\}$$

and

$$I_2 = \left\{\left(\sum_{i=1}^{m} \mathbf{X}_i^T(\mathbf{I}+\mathbf{Z}_i\hat{\mathbf{D}}\mathbf{Z}_i')^{-1}\mathbf{X}_i\right)^{-1} - \left(\sum_{i=1}^{m} \mathbf{X}_i^T(\mathbf{I}+\mathbf{Z}_i\mathbf{D}\mathbf{Z}_i')^{-1}\mathbf{X}_i\right)^{-1}\right\} \times \sum_{i=1}^{m} \mathbf{X}_i^T(\mathbf{I}+\mathbf{Z}_i\mathbf{D}\mathbf{Z}_i')^{-1}(\mathbf{Z}_i\mathbf{b}_i+\epsilon_i).$$

First notice that since $\mathbf{b}_i \sim N(0, D\sigma^2)$, $\epsilon_i \sim N(0, I_{n_i})$, by Central Limit Theorem and the regularity condition (C), we can show that

$$\sum_{i=1}^{m} \mathbf{X}_i^T\mathbf{Z}_i\mathbf{Z}_i'(\mathbf{Z}_i\mathbf{b}_i+\epsilon_i) = O_p(\sqrt{n})$$

and

$$\sum_{i=1}^{m} \mathbf{X}_i^T(\mathbf{I}+\mathbf{Z}_i\mathbf{D}\mathbf{Z}_i')^{-1}(\mathbf{Z}_i\mathbf{b}_i+\epsilon_i) = O_p(\sqrt{n}).$$

Now let's consider I_1. Because $\hat{\mathbf{D}} - \mathbf{D} = O_p(1/\sqrt{n})$,

$$\begin{aligned} &\sum_{i=1}^{m} \mathbf{X}_i^T(\mathbf{I}+\mathbf{Z}_i\hat{\mathbf{D}}\mathbf{Z}_i')^{-1}(\mathbf{Z}_i\mathbf{b}_i+\epsilon_i) - \sum_{i=1}^{m} \mathbf{X}_i^T(\mathbf{I}+\mathbf{Z}_i\mathbf{D}\mathbf{Z}_i')^{-1}(\mathbf{Z}_i\mathbf{b}_i+\epsilon_i) \\ = \;& O_p(1/\sqrt{n})\left|\sum_{i=1}^{m} \mathbf{X}_i^T\mathbf{Z}_i\mathbf{Z}_i'(\mathbf{Z}_i\mathbf{b}_i+\epsilon_i)\right| \qquad (4.20) \\ = \;& O_p(1/\sqrt{n})O_p(\sqrt{n}) = O_p(1) \qquad (4.21) \end{aligned}$$

Next because the elements of $\mathbf{Z}_i$ are bounded and $\hat{\mathbf{D}} - \mathbf{D} = O_p(1/\sqrt{n})$, it is not difficult to show that

$$\sum_{i=1}^{m} \mathbf{X}_i^T\mathbf{X}_i \le \sum_{i=1}^{m} \mathbf{X}_i^T(\mathbf{I}+\mathbf{Z}_i\hat{\mathbf{D}}\mathbf{Z}_i')^{-1}\mathbf{X}_i \le (1+C)\sum_{i=1}^{m} \mathbf{X}_i^T\mathbf{X}_i$$

where C is a positive constant determined by $\mathbf{Z}_i, i = 1, \ldots, m$ and $\mathbf{D}$. So

$$\begin{aligned} O_p(1/n) &= \left\{ \sum_{i=1}^{m} \mathbf{X}_i^T \mathbf{X}_i \right\}^{-1} \geq \left\{ \sum_{i=1}^{m} \mathbf{X}_i^T (\mathbf{I} + \mathbf{Z}_i \hat{\mathbf{D}} \mathbf{Z}_i') \mathbf{X}_i \right\}^{-1} \\ &\geq \frac{1}{1+C} \left\{ \sum_{i=1}^{m} \mathbf{X}_i^T \mathbf{X}_i \right\}^{-1} = O_p(1/n). \end{aligned} \tag{4.22}$$

By (4.20) and(4.22), we have that

$$I_1 = O_p(1/n) = o_p(1/\sqrt{n}). \tag{4.23}$$

Next consider I_2. First notice that

$$\begin{aligned} & \sum_{i=1}^{m} \mathbf{X}_i^T (\mathbf{I} + \mathbf{Z}_i \hat{\mathbf{D}} \mathbf{Z}_i')^{-1} \mathbf{X}_i - \sum_{i=1}^{m} \mathbf{X}_i^T (\mathbf{I} + \mathbf{Z}_i \mathbf{D} \mathbf{Z}_i')^{-1} \mathbf{X}_i \\ = \; & O_p(1/\sqrt{n}) \sum_{i=1}^{m} \left| \mathbf{X}_i^T \mathbf{Z}_i \mathbf{Z}_i' \mathbf{X}_i \right| \\ = \; & O_p(1/\sqrt{n}) O_p(n) \\ = \; & O_p(\sqrt{n}). \end{aligned}$$

and

$$\sum_{i=1}^{m} \mathbf{X}_i^T (\mathbf{I} + \mathbf{Z}_i \mathbf{D} \mathbf{Z}_i')^{-1} \mathbf{X}_i = O_p(n), \quad \sum_{i=1}^{m} \mathbf{X}_i^T (\mathbf{I} + \mathbf{Z}_i \hat{\mathbf{D}} \mathbf{Z}_i')^{-1} \mathbf{X}_i = O_p(n),$$

hence

$$I_2 = \frac{O_p(\sqrt{n})}{O_p(n) \cdot O_p(n)} \cdot O_p(\sqrt{n}) = O_p(\sqrt{n}/n) = o_p(1/\sqrt{n}) \tag{4.24}$$

Finally, by (4.23) and (4.24), (4.19) is easy to obtain and the proof of Proposition 4.2.1 is complete. □

Proof of Proposition 4.2.2: If $\boldsymbol{\beta}$ is known, then we have

$$\mathbf{u}_i = \mathbf{Y}_i - \mathbf{X}_i \boldsymbol{\beta} \quad \text{and} \quad \tilde{\mathbf{b}}_i = (\mathbf{Z}_i^T \mathbf{Z}_i)^{-1} \mathbf{Z}_i^T \mathbf{u}_i.$$

Then an estimate of $\mathbf{D}$ is given

$$\tilde{\mathbf{D}} = \frac{1}{m\hat{\sigma}^2}\sum_{i=1}^{m}\tilde{\mathbf{b}}_i\tilde{\mathbf{b}}_i^T - \frac{1}{m}\sum_{i=1}^{m}(\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}. \tag{4.25}$$

where $\hat{\sigma}^2$ is an estimate of σ^2 defined by the following (4.26).

Let $\hat{\boldsymbol{\beta}}$ is the $\sqrt{n}$ consistent estimate of $\boldsymbol{\beta}$, we can define

$$\hat{\mathbf{u}}_i = \mathbf{Y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}} \quad \text{and} \quad \hat{\mathbf{b}}_i = (\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}\mathbf{Z}_i^T\hat{\mathbf{u}}_i,$$

and an estimate of σ^2 can be defined as

$$\hat{\sigma}^2 = (n-qm)^{-1}\sum_{i=1}^{m}\mathrm{RSS}_i, \quad \mathrm{RSS}_i = \hat{\mathbf{u}}_i^T(\mathbf{I}_{n_i} - \mathbf{P}_i)\hat{\mathbf{u}}_i. \tag{4.26}$$

By Lemma 4.2.1, it is known that $\hat{\sigma}^2$ is a $\sqrt{n}$-consistent estimate of σ^2. Then $\mathbf{D}$ can be estimated as

$$\hat{\mathbf{D}} = \frac{1}{m\hat{\sigma}^2}\sum_{i=1}^{m}\hat{\mathbf{b}}_i\hat{\mathbf{b}}_i^T - \frac{1}{m}\sum_{i=1}^{m}(\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}. \tag{4.27}$$

Next we first prove that

$$\tilde{\mathbf{D}} - \hat{\mathbf{D}} = o_p(1/\sqrt{n}).$$

Then we only need study the asymptotic distribution of $\sqrt{n}\tilde{\mathbf{D}}$. By (4.25) and (4.38), we have

$$\tilde{\mathbf{D}} - \hat{\mathbf{D}} = \frac{1}{m\hat{\sigma}^2}\sum_{i=1}^{m}\tilde{\mathbf{b}}_i\tilde{\mathbf{b}}_i^T - \frac{1}{m\hat{\sigma}^2}\sum_{i=1}^{m}\hat{\mathbf{b}}_i\hat{\mathbf{b}}_i^T \tag{4.28}$$

Because

$$\begin{aligned}\hat{\mathbf{u}}_i &= \mathbf{Y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}} = \mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta} + \mathbf{X}_i\boldsymbol{\beta} - \mathbf{X}_i\hat{\boldsymbol{\beta}} \\ &= \mathbf{u}_i + \mathbf{X}_i(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\end{aligned}$$

and

$$\begin{aligned}\hat{\mathbf{b}}_i &= (\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}\mathbf{Z}_i^T\hat{\mathbf{u}}_i \\ &= (\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}\mathbf{Z}_i^T\mathbf{u}_i + (\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}\mathbf{Z}_i^T\mathbf{X}_i(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \\ &\hat{=} \tilde{\mathbf{b}}_i + \mathbf{e}_i,\end{aligned}$$

Then

$$\tilde{\mathbf{D}} - \hat{\mathbf{D}} = -\frac{1}{m\hat{\sigma}^2}\sum_{i=1}^{m}\{\tilde{\mathbf{b}}_i\mathbf{e}_i^T + \mathbf{e}_i\tilde{\mathbf{b}}_i^T\} - \frac{1}{m\hat{\sigma}^2}\sum_{i=1}^{m}\mathbf{e}_i\mathbf{e}_i^T.$$

On the other hand, it is easy to show that

$$\tilde{\mathbf{b}}_i = (\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}\mathbf{Z}_i^T\mathbf{u}_i = \mathbf{b}_i + (\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}\mathbf{Z}_i^T\epsilon_i.$$

Hence by the regularity conditions and the definition of $\mathbf{b}_i$, it can be shown that

$$\begin{aligned}
\frac{1}{m\hat{\sigma}^2}\sum_{i=1}^{m}\tilde{\mathbf{b}}_i\mathbf{e}_i^T &= \frac{1}{m\hat{\sigma}^2}\sum_{i=1}^{m}(\mathbf{b}_i + (\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}\mathbf{Z}_i^T\epsilon_i)(\boldsymbol{\beta}^T - \hat{\boldsymbol{\beta}}^T)\mathbf{X}_i^T\mathbf{Z}_i(\mathbf{Z}_i^T\mathbf{Z}_i)^{-1} \\
&\leq O_p(1/\sqrt{n})\left(\left|\frac{1}{m\hat{\sigma}^2}\sum_{i=1}^{m}\mathbf{b}_i\mathbf{X}_i^T\mathbf{Z}_i(\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}\right|\right. \\
&\quad \left. + \left|\frac{1}{m\hat{\sigma}^2}\sum_{i=1}^{m}(\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}\mathbf{Z}_i^T\epsilon_i\mathbf{X}_i^T\mathbf{Z}_i(\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}\right|\right) \qquad (4.29) \\
&= O_p(1/\sqrt{n})O_p(1/\sqrt{m}) = O_p(1/n). \qquad (4.30)
\end{aligned}$$

Similarly, we also have

$$\frac{1}{m\hat{\sigma}^2}\sum_{i=1}^{m}\mathbf{e}_i\tilde{\mathbf{b}}_i^T = O_p(1/n) \qquad (4.31)$$

It is also easy to find

$$\begin{aligned}
\frac{1}{m\sigma^2}\sum_{i=1}^{m}\mathbf{e}_i\mathbf{e}_i^T &= \frac{1}{m\sigma^2}\sum_{i=1}^{m}(\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}\mathbf{Z}_i^T\mathbf{X}_i(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T\mathbf{X}_i^T\mathbf{Z}_i(\mathbf{Z}_i^T\mathbf{Z}_i)^{-1} \\
&= O_p(1/n)\left|\frac{1}{m\sigma^2}\sum_{i=1}^{m}(\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}\mathbf{Z}_i^T\mathbf{X}_i\mathbf{X}_i^T\mathbf{Z}_i(\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}\right| \\
&= O_p(1/n)O_p(1) = O_p(1/n). \qquad (4.32)
\end{aligned}$$

So by (4.29), (4.31) and (4.32), it is obtained that

$$\tilde{\mathbf{D}} - \hat{\mathbf{D}} = O_p(1/n). \qquad (4.33)$$

For $\tilde{\mathbf{D}}$, this can be written as

$$\tilde{\mathbf{D}} = \frac{1}{m\sigma^2}\sum_{i=1}^{n}\tilde{\mathbf{b}}_i\tilde{\mathbf{b}}_i^T - \frac{1}{m}\sum_{i=1}^{n}(\mathbf{Z}_i^T\mathbf{Z}_i)^{-1} + \left\{\frac{1}{m\hat{\sigma}^2} - \frac{1}{m\sigma^2}\right\}\sum_{i=1}^{n}\tilde{\mathbf{b}}_i\tilde{\mathbf{b}}_i^T \hat{=} D_1 + D_2.$$

Notice from the definition of $\hat{\sigma}^2$, it is not difficult to show that D_1 and D_2 are linear independent and $\mathrm{E}D_2 \to 0$. Hence

$$\mathrm{Var}(\mathrm{vec}(\tilde{\mathbf{D}})) = \mathrm{Var}(\mathrm{vec}(D_1)) + \mathrm{Var}(\mathrm{vec}(D_2))$$

and

$$\mathrm{E}(\tilde{\mathbf{D}}) \to \mathrm{E}(D_1) \quad as \quad n \to \infty.$$

For D_1, we have

$$\begin{aligned} D_1 &= \frac{1}{m\sigma^2}\sum_{i=1}^{m}\mathbf{b}_i\mathbf{b}_i^T + \left\{\frac{1}{m\sigma^2}\sum_{i=1}^{m}(\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}\mathbf{Z}_i^T\epsilon\epsilon_i^T\mathbf{Z}_i(\mathbf{Z}_i^T\mathbf{Z}_i)^{-1} - \frac{1}{m}\sum_{i=1}^{m}(\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}\right\} \\ &+ \frac{1}{m\sigma^2}\sum_{i=1}^{m}\left\{(\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}\mathbf{Z}_i^T\epsilon_i\mathbf{b}_i^T + \mathbf{b}_i\epsilon_i^T\mathbf{Z}_i(\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}\right\} \\ &\hat{=} D_{11} + D_{12} + D_{13} \end{aligned}$$

It is easy to see that D_{11}, D_{12} and D_{13} are independent and $\mathrm{E}D_{12} = \mathrm{E}D_{13} = 0$, so we have

$$\mathrm{E}D_1 = \mathrm{E}D_{11} \quad \text{and} \quad \mathrm{Var}(\mathrm{vec}(D_1)) = \mathrm{Var}(\mathrm{vec}(D_{11})) + \mathrm{Var}(\mathrm{vec}(D_{12})) + \mathrm{Var}(\mathrm{vec}(D_{13}))$$

It is obvious that

$$\mathrm{E}D_1 = \mathbf{D} \quad \text{and} \quad \mathrm{Var}(\sqrt{m}\cdot\mathrm{vec}(D_{11})) = \frac{1}{\sigma^4}\mathrm{E}\{\mathbf{b}_1\mathbf{b}_1^T \otimes \mathbf{b}_1\mathbf{b}_1^T\} - \mathbf{D}\otimes\mathbf{D} \qquad (4.34)$$

D_{12} can be written as

$$\begin{aligned} &\frac{1}{m\sigma^2}\sum_{i=1}^{m}(\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}\mathbf{Z}_i^T\epsilon_i^T\epsilon_i\mathbf{Z}_i(\mathbf{Z}_i^T\mathbf{Z}_i)^{-1} - \frac{1}{m}\sum_{i=1}^{m}(\mathbf{Z}_i^T\mathbf{Z}_i)^{-1} \\ =\ &\frac{1}{m\sigma^2}\sum_{i=1}^{m}(\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}\mathbf{Z}_i^T(\epsilon_i^T\epsilon_i - \sigma^2\mathbf{I}_{n_i})\mathbf{Z}_i(\mathbf{Z}_i^T\mathbf{Z}_i)^{-1} \\ =\ &\frac{1}{m\sigma^2}\sum_{i=1}^{m}(\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}\mathbf{Z}_i^T\mathbf{A}\mathbf{Z}_i(\mathbf{Z}_i^T\mathbf{Z}_i)^{-1} + \frac{1}{m\sigma^2}\sum_{i=1}^{m}(\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}\mathbf{Z}_i^T\mathbf{B}\mathbf{Z}_i(\mathbf{Z}_i^T\mathbf{Z}_i)^{-1} \\ \hat{=}\ &D_{121} + D_{122} \end{aligned}$$

where

$$\mathbf{A} = \mathrm{diag}(\epsilon_{i1}^2 - \sigma^2, \ldots, \epsilon_{in_i}^2 - \sigma^2)$$

and

$$B_{jl} = \epsilon_{ij}\epsilon_{il}, j \neq l \quad \text{and} \quad B_{jj} = 0, j, l = 1, \ldots, n_i.$$

It is not difficult to show that

$$\mathrm{Var}(\sqrt{m} \cdot \mathrm{vec}(D_{12})) = \mathrm{Var}(\sqrt{m} \cdot \mathrm{vec}(D_{121})) + \mathrm{Var}(\sqrt{m} \cdot \mathrm{vec}(D_{122}))$$

and by some complex calculation,

$$\begin{aligned} \mathrm{Var}(\sqrt{m} \cdot \mathrm{vec}(D_{121})) &= \mathrm{Var}(\epsilon_{11}^2)\Delta_3/\sigma^4, & (4.35) \\ \mathrm{Var}(\sqrt{m} \cdot \mathrm{vec}(D_{122})) &= 2(\Delta_2 - \Delta_3). & (4.36) \end{aligned}$$

For D_{13}, we have

$$\mathrm{Var}(\sqrt{m} \cdot \mathrm{vec}(D_{13})) = \frac{1}{\sigma^2}\{\mathbf{D} \otimes \Gamma + \Gamma \times \mathbf{D} + \Delta_1\} \tag{4.37}$$

Next consider D_2. By Lemma (2.1), $\hat{\sigma}^2 - \sigma^2 = O_p(1/\sqrt{n})$, hence

$$D_2 = \Big\{\frac{1}{m\hat{\sigma}^2} - \frac{1}{m\sigma^2}\Big\} \sum_{i=1}^{n} \tilde{\mathbf{b}}_i \tilde{\mathbf{b}}_i^T = \frac{\hat{\sigma}^2 - \sigma^2}{\sigma^2}\Big(\mathbf{D} + \frac{1}{m}\sum_{i=1}^{m} (\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}\Big) + o_p(1/\sqrt{n}),$$

and

$$\begin{aligned} & \mathrm{Var}(\sqrt{m} \cdot \mathrm{vec}(D_2)) \\ = \; & \big(2(1+\gamma)c_1 + \mathrm{Var}(\epsilon_{11}^2)\gamma c_1\big)\, \mathrm{vec}\left\{\mathbf{D} + \frac{1}{m}\sum_{i=1}^{m}(\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}\right\} \mathrm{vec}\left\{\mathbf{D} + \frac{1}{m}\sum_{i=1}^{m}(\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}\right\}^T \\ = \; & \big(2(1+\gamma)c_1 + \mathrm{Var}(\epsilon_{11}^2)\gamma c_1\big)\, \Delta_4 \qquad (4.38) \end{aligned}$$

Finally, by (4.33)—(4.38), the transformation from *vec* to *vech* by R_q, and the central limited theory, the proof of Proposition 4.2.2 is complete. □

Proof of Theorem 4.3.1: Define

$$\begin{aligned} Q(\boldsymbol{\beta}) &= \sum_{i=1}^{m}(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta})^T(\mathbf{I} + \mathbf{Z}_i\hat{\mathbf{D}}\mathbf{Z}_i')^{-1}(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta}) + n\sum_{j=1}^{p} p_{\lambda_n}(|\beta_j|) \\ &= L(\boldsymbol{\beta}) + n\sum_{j=1}^{p} p_{\lambda_n}(|\beta_j|) \end{aligned}$$

and $\alpha_n = (1/\sqrt{n})$.

To prove the theorem, we first show that for any give $\varepsilon > 0$, there exist a large constant C such that

$$\mathrm{P}\left\{\sup_{\|\mathbf{u}\|=C} Q(\boldsymbol{\beta}_0 + \alpha_n\mathbf{u}) > Q(\boldsymbol{\beta}_0)\right\} > 1 - \varepsilon \tag{4.39}$$

This implies with probability at least $1-\varepsilon$ that there exist local minimizer in the ball $\{\beta_0 + \alpha_n\mathbf{u} : \|\mathbf{u}\| \leq C\}$. Hence there exists a local minimizer such that $\|\boldsymbol{\beta}^* - \boldsymbol{\beta}_0\| = O_p(\alpha_n)$.

Since $p_{\lambda_n}(0) = 0$, we have

$$D_n(\mathbf{u}) = Q(\boldsymbol{\beta}_0+\alpha_n\mathbf{u})-Q(\boldsymbol{\beta}_0) \geq L(\boldsymbol{\beta}_0+\alpha_n\mathbf{u})-L(\boldsymbol{\beta}_0)+n\sum_{j=1}^{s}\left\{p_{\lambda_n}(|\boldsymbol{\beta}_{j0} + \alpha_n u_j|) - p_{\lambda_n}(|\beta_{j0}|)\right\}$$

where s is the number of components of $\boldsymbol{\beta}_{10}$. By simple calculation, it is not difficult to show that

$$\begin{aligned} D_n(\mathbf{u}) &\geq \alpha_n^2\sum_{i=1}^{m}\mathbf{u}^T\mathbf{X}_i^T(\mathbf{I} + \mathbf{Z}_i\hat{\mathbf{D}}\mathbf{Z}_i^T)^{-1}\mathbf{X}_i\mathbf{u} - \alpha_n\sum_{i=1}^{m}(\mathbf{Z}_i\mathbf{b}_i + \epsilon_i)^T(\mathbf{I} + \mathbf{Z}_i\hat{\mathbf{D}}\mathbf{Z}_i^T)^{-1}\mathbf{X}_i\mathbf{u} \\ &- \alpha_n\sum_{i=1}^{m}\mathbf{u}^T\mathbf{X}_i^T(\mathbf{I} + \mathbf{Z}_i\hat{\mathbf{D}}\mathbf{Z}_i^T)^{-1}(\mathbf{Z}_i\mathbf{b}_i + \epsilon_i) \\ &+ n\sum_{j=1}^{s}\left\{p_{\lambda_n}(|\boldsymbol{\beta}_{j0} + \alpha_n u_j|) - p_{\lambda_n}(|\beta_{j0}|)\right\} \\ &\hat{=} D_{n1} + D_{n2} + D_{n3} + D_{n4}. \end{aligned}$$

Under the regularity condition (C), we have

$$D_{n1} \geq O_p(n\alpha_n^2)\|\mathbf{u}\|^2, \tag{4.40}$$

and

$$D_{n2} = O_p(\alpha_n\sqrt{n})\|\mathbf{u}\| \quad \text{and} \quad D_{n3} = O_p(\alpha_n\sqrt{n})\|\mathbf{u}\|. \tag{4.41}$$

After taking the second order Taylor expansion of the first term around β_{j0} in D_{n4} we have

$$D_{n4} = n\sum_{j=1}^{s} p'_{\lambda_n}(|\beta_{j0}|)\text{sgn}(\beta_{j0})\alpha_n u_j + n\sum_{j=1}^{s} p''_{\lambda_n}(|\beta_{j0}|)\alpha_n^2 u_j^2 \cdot (1+o(1))$$

By the definition of the SCAD function, as $\lambda_n \to 0$, $a_n = \max\{p'_{\lambda_n}(|\beta_{j0}|) : \beta_{j0} \neq 0\} = 0$ and $\max(p''_{\lambda_n}(|\beta_{j0}|)) \to 0$. Hence when n is large enough, $D_{n4} = 0$. Moreover it is obvious that D_{n1} dominates D_{n2} and D_{n3}, therefore (4.39) holds. In other words, there is a local minimizer $\hat{\boldsymbol{\beta}}$ of (4.17) such that

$$\|\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}\| = O_p(\alpha_n).$$

Next we show this minimizer $\hat{\boldsymbol{\beta}}$ has properties of (a) and (b). In fact if (a) is true, by the oracle properties of SCAD penalty, we know that the asymptotic normality of $\hat{\boldsymbol{\beta}}$ can be directly deduced from Proposition 4.2.1. Hence we only need to show that $\hat{\boldsymbol{\beta}}$ has the property (a).

For $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| = O(1/\sqrt{n})$ and $\beta_{j0} = 0$,$j = s+1, \ldots, p$, we consider the derivative of $Q(\boldsymbol{\beta})$ with respect to β_j,

$$\begin{aligned}\frac{\partial Q(\boldsymbol{\beta})}{\beta_j} &= -\sum_{i=1}^{m} 2\mathbf{X}_{ij}^T(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta})(\mathbf{I} + \mathbf{Z}_i\hat{\mathbf{D}}\mathbf{Z}_i^T)^{-1} + np'_{\lambda_n}(|\beta_j|)\text{sgn}(\beta_j) \\ &= Q_1 + Q_2.\end{aligned}$$

Based on the definition of SCAD penalty function, $p'_{\lambda_n}(|\beta_j|) = \lambda_n$ when $\beta_j = o(1/\sqrt{n})$ and $\sqrt{n}\lambda_n \to \infty$. Hence $Q_2 = n\lambda_n\text{sgn}(\beta_j)$ when n is large enough.

Under the regularity condition (C), we know that

$$\sum_{i=1}^{m} 2\mathbf{X}_{ij}^T(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta})(\mathbf{I} + \mathbf{Z}_i\hat{\mathbf{D}}\mathbf{Z}_i^T)^{-1} = O_p(\sqrt{n}).$$

Since $\sqrt{n}\lambda_n \to \infty$ when $n \to \infty$, Q_1 is dominated by Q_2 and Q_2 determines the sign of the derivative above. This means that for some small $\varepsilon_n = Cn^{-1/2}$ and $j = s+1, \ldots, p$, we have

$$\frac{\partial Q(\boldsymbol{\beta})}{\beta_j} > 0 \qquad \text{for} \quad 0 < \beta_j < \varepsilon_n \tag{4.42}$$

$$\frac{\partial Q(\boldsymbol{\beta})}{\beta_j} < 0 \qquad \text{for} \quad 0 > \beta_j > -\varepsilon_n \tag{4.43}$$

Therefore, only when $\beta_j = 0, j = s+1, \ldots, p$, $Q(\boldsymbol{\beta})$ arrives its minimizer point.

Hereby we finish the proof of this theorem. □

Proof of Theorem 4.3.2: Similar as the proof of Proposition 4.2.2, we only need to show there exists a local minimizer $\mathbf{D}^*$ such that

$$\|\mathbf{D} - \mathbf{D}^*\| = O_p(\sqrt{\log n/n}), \quad \text{and} \quad \mathbf{d}_2^* = 0.$$

The asymptotic normality of $\mathbf{D}_1^*$ follows by the properties of the SCAD penalty function.

To show $\|D - D^*\| = O_p(\sqrt{\log n/n})$, it suffices to show $\|\hat{D} - D^*\| = O_p(\sqrt{\log n/n})$ since we showed in Proposition 4.2.2 that $\|\hat{D} - D\| = O_p(\sqrt{1/n})$. Moreover, since

$$D^* = \frac{1}{m}\sum_{i=1}^{m} b_i^* b_i^{*T} - \frac{1}{m}\sum_{i=1}^{m} \mathbf{Z}_i\mathbf{Z}_i^T$$

and

$$\hat{D} = \frac{1}{m}\sum_{i=1}^{m} \hat{b}_i \hat{b}_i^T - \frac{1}{m}\sum_{i=1}^{m} \mathbf{Z}_i\mathbf{Z}_i^T$$

we only need to show $\|\hat{\mathbf{B}} - \mathbf{B}^*\| = O_p(\sqrt{\log(n)/n})$, where $\hat{\mathbf{B}} = (\hat{\mathbf{b}}_1^T, \ldots, \hat{\mathbf{b}}_m^T)^T$.

First define $\hat{\mathbf{u}}_i = \mathbf{Y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}$ and

$$Q(\mathbf{B}) = \sum_{i=1}^{m}(\hat{\mathbf{u}}_i - \mathbf{Z_i}\mathbf{b}_i)^T(\hat{\mathbf{u}}_i - \mathbf{Z_i}\mathbf{b}_i) + \sum_{i=1}^{q} np_{\xi_n}(c_k),$$

where

$$c_k = \left| \frac{1}{m\hat{\sigma}^2}\sum_{i=1}^{m} b_{ik}^2 - \frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{n_i} Z_{ijk}^2 \right|^{\frac{1}{2}}.$$

To prove the theorem, we need to show that for any give $\varepsilon > 0$, there exists a large constant C such that

$$\mathrm{P}\left\{ \sup_{\|\mathbf{v}\|=C} Q(\hat{\mathbf{B}} + \alpha_n\mathbf{v}) > Q(\hat{\mathbf{B}}) \right\} > 1 - \varepsilon \tag{4.44}$$

where $\alpha_n = \sqrt{\log n/n}$, $\hat{\mathbf{B}} = (\hat{\mathbf{b}}_1^T, \ldots, \hat{\mathbf{b}}_m^T)^T$ and $\hat{\mathbf{b}}_i = (\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}\mathbf{Z}_i^T\hat{\mathbf{u}}_i$ is defined as in the proof of Proposition 4.2.2. This implies with probability at least $1-\varepsilon$ that there exists a local minimizer such that $\|\mathbf{B}^* - \hat{\mathbf{B}}\| = O_p(\sqrt{\log n/n})$ and $\|\hat{\mathbf{D}} - \mathbf{D}^*\| = O_p(\sqrt{\log n/n})$. Then it is easy to see that by Theorem 2.2, $\|\mathbf{D} - \mathbf{D}^*\| = O_p(\sqrt{\log n/n})$.

By the definition of $\hat{\mathbf{B}}$, we have

$$\begin{aligned} Q(\hat{\mathbf{B}} + \alpha_n\mathbf{u}) - Q(\hat{\mathbf{B}}) &= \sum_{i=1}^{m}\Big((\hat{\mathbf{u}}_i - \mathbf{Z}_i\hat{\mathbf{b}}_i)^T\mathbf{Z}_i\alpha_n v_i + \alpha v_i^T\mathbf{Z}_i^T(\hat{\mathbf{u}}_i - \mathbf{Z}_i\hat{\mathbf{b}}_i)\mathbf{Z}_i \\ &\quad + \alpha_n^2 v_i^T\mathbf{Z}_i^T\mathbf{Z}_i v_i\Big) + \sum_{i=1}^{q} n(p_{\xi_n}(\tilde{c}_k) - p_{\xi_n}(\hat{c}_k)) \\ &= Q_1 + Q_2 \end{aligned}$$

where $\hat{c}_k = \left| \frac{1}{m\hat{\sigma}^2}\sum_{i=1}^{m}\hat{b}_{ik}^2 - \frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{n_i} Z_{ijk}^2 \right|^{\frac{1}{2}}$ and $\tilde{c}_k = \left| \frac{1}{m\hat{\sigma}^2}\sum_{i=1}^{m}(\hat{b}_{ik} + \alpha_n v_i)^2 - \frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{n_i} Z_{ijk}^2 \right|^{\frac{1}{2}}$.

First by the definitions of $\hat{\mathbf{b}}_i$, we know that the first two terms in Q_1 are equal to 0. By the regularity conditions, the third term of Q_1 is of order $O_p(n\alpha_n^2)C^2$.

After taking the Taylor expansion of $p_{\xi_n}(\tilde{c}_k)$ around $\hat{c}_k$,

$$Q_2 = \sum_{i=1}^{q} np'_{\xi_n}(\hat{c}_k)(\tilde{c}_k - \hat{c}_k) + \sum_{i=1}^{q} np''_{\xi_n}(\hat{c}_k)(\tilde{c}_k - \hat{c}_k)^2(1 + o(1))$$

For Q_2, because $\sqrt{n/\log n}\cdot\xi_n \to \infty$ and Proposition 4.2.2, when n is large enough, we have $p'_{\xi_n}(\hat{c}_k)$ is bounded by ξ_n and

$$\max\{p''_{\xi_n}(\hat{c}_k), k = 1, \ldots, q\} \to 0.$$

On the other hand, since $\frac{1}{m}\sum_{i=1}^{m} \hat{\mathbf{b}}_i = O_p(1/\sqrt{n})$, by regularity conditions we know that

$$\begin{aligned}
& \frac{1}{m\tilde{\sigma^2}} \sum_{i=1}^{m} (\hat{\mathbf{b}}_i + \alpha_n \mathbf{u}_i)^T (\hat{\mathbf{b}}_i + \alpha_n \mathbf{u}_i) - \frac{1}{m\hat{\sigma^2}} \sum_{i=1}^{m} \hat{\mathbf{b}}_i \hat{\mathbf{b}}_i^T \\
= & \frac{1}{m\hat{\sigma^2}} \sum_{i=1}^{m} \left(\alpha_n \hat{\mathbf{b}}_i^T v_i + \alpha_n v_i^T \hat{\mathbf{b}}_i + \alpha_n^2 v_i^T v_i \right) + \left(\frac{1}{m\hat{\sigma^2}} - \frac{1}{m\hat{\sigma^2}}\right) \sum_{i=1}^{m} \hat{\mathbf{b}}_i \hat{\mathbf{b}}_i^T \\
= & O_p(\alpha_n^2) \cdot C^2
\end{aligned}$$

Hence $\tilde{c}_k^2 - \hat{c}_k^2 = O_p(\alpha_n^2) \cdot C^2$, and $\tilde{c}_k - \hat{c}_k \leq O_p(\alpha_n) \cdot C$.

$$Q_2 = O_p(n\xi_n\alpha_n) \cdot C + o_p(n\alpha_n^2) \cdot C^2 = O_p(n\alpha_n^2) \cdot C + o_p(n\alpha_n^2) \cdot C^2.$$

It is obvious that Q_2 is dominated by Q_1 when C is large enough and $Q(\hat{\mathbf{B}} + \alpha_n \mathbf{u}) - Q(\hat{\mathbf{B}}) > 0$. Hence it is easy to see that (4.44) has been proved.

Next we want to show that $\mathbf{d}_2^* = 0$. To simplify the analysis, assume that $D_{qq} = 0$, and $\mathbf{D}^* - \mathbf{D}_0 = O_p(\sqrt{\log n/n})$. We want to show that $c_q^* = 0$ where c_q^* is the estimate of $D_{qq}^{\frac{1}{2}}$. First we assume that $c_q^* = O_p(\sqrt{\log n/n}) \neq 0$. For $\quad i = 1, \ldots, q$,

$$\begin{aligned}
\frac{\partial Q(\mathbf{B}^*)}{\partial b_{iq}} &= (\hat{\mathbf{u}}_i - \mathbf{Z}_i \mathbf{b}_i^*)^T \mathbf{Z}_{iq} + n p'_{\xi_n}(c_q^*) \cdot \frac{b_{iq}^* \text{sign}(c_q^*)}{m c_q^* \hat{\sigma}^2} \\
&= (\mathbf{Z}_i (\mathbf{Z}_i^T \mathbf{Z}_i)^{-1} \mathbf{Z}_i^T \hat{\mathbf{u}}_i - \mathbf{Z}_i \mathbf{b}_i^*)^T \mathbf{Z}_{iq} + n p'_{\xi_n}(|c_q^*|) \cdot \frac{b_{iq}^* \text{sign}(c_q^*)}{m c_q^* \hat{\sigma}^2} \\
&= (\mathbf{Z}_i \hat{\mathbf{b}}_i - \mathbf{Z}_i \mathbf{b}_i^*)^T \mathbf{Z}_{iq} + n p'_{\xi_n}(|c_q^*|) \cdot \frac{b_{iq}^* \text{sign}(c_q^*)}{m c_q^* \hat{\sigma}^2}
\end{aligned}$$

Furthermore,

$$\begin{aligned}
\sum_{i=1}^{m} \frac{\partial Q(\mathbf{B}^*)}{\partial b_{iq}} \cdot b_{iq}^* &= \sum_{i=1}^{m} (\mathbf{Z}_i \hat{\mathbf{b}}_i - \mathbf{Z}_i \mathbf{b}_i^*)^T \mathbf{Z}_{iq} b_{iq}^* + n p'_{\xi_n}(|c_q^*|) \cdot \sum_{i=1}^{m} \frac{b_{iq}^{*2} \text{sign}(c_q^*)}{m c_q^* \hat{\sigma^2}} \\
&\hat{=} Q_{d1} + Q_{d2}.
\end{aligned}$$

For Q_{d1}, by Cauchy inequality, n_i and Z_{ijq} are bounded by constants, therefore

$$\begin{aligned} Q_{d1}^2 &\leq \left\{\sum_{i=1}^{m} \|\mathbf{b}_i - \hat{\mathbf{b}}_i\|^2\right\}\left\{\sum_{i=1}^{m} \|\mathbf{Z}_i^T\mathbf{Z}_{iq}\|^2 b_{iq}^2\right\} = O_p(\log n)\cdot\sum_{i=1}^{m} b_{iq}^2 \\ &= O_p(n\log n)\cdot\left\{\frac{1}{m\hat{\sigma}^2}\sum_{i=1}^{m} b_{iq}^2 - \frac{1}{m}\sum_{i=1}^{m}\mathbf{Z}_{iq}^T\mathbf{Z}_{iq}\right\} + O_p(n\log n)\cdot\left\{\frac{1}{m}\sum_{i=1}^{m}\mathbf{Z}_{iq}^T\mathbf{Z}_{iq}\right\} \\ &= O_p(n\log(n))\cdot O_p(\sqrt{\log(n)/n}) + O_p(n\log n) \qquad (4.45) \\ &= O_p(n\log n). \qquad (4.46) \end{aligned}$$

For Q_{d2}, because $\mathbf{D}^* - \mathbf{D}_0 = O_p(\sqrt{\log n/n})$, we know that $c_k^* = O_p(\sqrt{\log n/n})$, and hence

$$\begin{aligned} Q_{d2} &= np'_{\xi_n}(|c_q|)\cdot\frac{\text{sign}(c_q)}{mc_q}\left\{\sum_{i=1}^{m}\frac{b_{iq}^2}{\hat{\sigma}^2} - \sum_{i=1}^{m}\mathbf{Z}_{iq}^T\mathbf{Z}_{iq} + \sum_{i=1}^{m}\mathbf{Z}_{iq}^T\mathbf{Z}_{iq}\right\} \\ &= np'_{\xi_n}(|c_q|)\cdot\text{sign}(c_q)c_q + np'_{\xi_n}(|c_q|)\cdot\frac{\text{sign}(c_q)}{c_q}\cdot O(1) \\ &= O_p(\sqrt{n\log n}\cdot\sqrt{\log n/n}) + O(\sqrt{n\log n}\cdot\sqrt{n/\log n})\cdot\text{sign}(c_p) \qquad (4.47) \end{aligned}$$

If $\mathbf{B}^* = (\mathbf{b}_1^{*T},\ldots,\mathbf{b}_m^{*T})^T$ is the minimizer point of $Q(\mathbf{B})$ and c_q^* does not equal to zero, we should have that

$$Q_{1d} + Q_{2d} = 0,$$

However, it can be easily seen that Q_{d1} is dominated by Q_{d2} and the sign of c_q determines the sign of $Q_{1d} + Q_{2d}$, which cannot equal to zero. This contradicts the assumption that $c_q^* = O_p(\sqrt{\log n/n}) \neq 0$. Hence it is a necessary condition that $c_q^* = 0$ if $c_q^* = O_p(\sqrt{\log n/n})$ is also a local minimizer. Therefore for the local minimizer $\mathbf{b}_i^*, i = 1,\ldots,m$, the sparsity property must hold. The proof of this Theorem has been finished. □

Bibliography

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, in: B.N. Petrov, F. Csáki(Eds.), Second Internal Symposium on Information Theory. Akadémiai Kiado, Budapest, 1973, 267-281.

Anker, M. (2003). *Investigating Cause of Death During an Outbreak of Ebola Virus Haemorrhagic Fever: Draft Verbal Autopsy Instrument.* World Health Organization, Geneva.

Antoniadis, A., Fan, J. (2001). Regularization of Wavelet Approximations: Rejoinder. *Journal of the American Statistical Association*, **96**, 964967.

Bafumi, J., Gelman, A., Park, D. K. and Kaplan, N.(2005). Practical issues in implementing and understanding Bayesian ideal point estimation. *Political Analysis*, **14**, 381-396.

Boulle, A., Chandramohan, D. and Weller, P. (2001). A case study of using artificial neural networks for classifying cause of death from verbal autopsy. *International Journal of Epidemiology*, **30**, 515520.

Bradlow, E.T., Wainer, H., and Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, **64**, 153-168.

Breiman,L.(1996). Heuristics of Instability and Stablilization in Model Selection, *Annals of Statistics*, **24**, 2350-2383.

Bryk and Raudenbush (2001) Hierarchical Linear Models: Applications and Data Analysis Methods, 2nd ed. Sage Publication.

Chandramohan, D., Maude, G. H., Rodrigues, L. C. and Hayes, R. J. (1994). Verbal autopsies for adult deaths: Issues in their development and validation. *International Journal of Epidemiology*, **23**, 213222.

Chandramohan, D., Setel, P. and Quigley, M. (2001). Effect of misclassification of causes of death in verbal autopsy: Can it be adjusted. *International Journal of Epidemiology*, **30**, 509514.

Chang, C.-C. and Lin, C.-J. (2001). LIBSVM: A library for support vector machines. Available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

Chen, Z. and Dunson D. B. (2003). Random effects selection in linear mixed models.

Biometrics, **59**, 762-769.

Clinton, J., Jackman, S. and Rivers, D.(2004). The statistical analysis of roll call data. *American Political Science Review*, **98**, 355-370.

Dawes, R. M., Faust, D. and Meehl, P. E. (1989). Clinical versus actuarial judgement. *Science*, **243**, 16681674.

Enelow, J. and Hinich M. (1984). *The Spatial Theory of Voting: An Introduction*. Cambridge: Cambridge University Press.

Epstein, L. and Mershon, C. (1996). Measuring political preferences. *American Journal of Political Science*, **60**, 801-818.

Epstein, L., Segal, J.A. , Spaeth, H.J, and Walker, T.G.(2003). *The Supreme Court Compendium: Data, Decisions and Development*. 2nd ed. Congressional Quarterly Inc. Washington, D.C..

Fan,J.(1997). Comments on "Wavelets in Statistics: A review," by A.Antoniadis. *J. Italian Statist. Soc.* 6, 131-138.

Fan, J., Peng H. and Huang T. (2005). Semilinear high-dimensional model for normalization of microarray data: a theoretical analysis and partial consistency, with discussion. *Journal of the American Statistical Association*, **100**, 781-813.

Fan,J., and Li,R. (2001). Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties, *Journal of the American Statistical Association*,96, **456**, 1348-1360.

Fan, J. and Li, R. (2004). New Estimation and Model Selection Procedures for Semiparametric Modeling in Longitudinal Data Analysis. *Journal of American Statistical Association*, **99**, 710-723.

Fan, J. and Peng H. (2004). Nonconcave penalized likelihood with a diverging number of parameters, *The annals of Statistics*, **32**, 928-961.

Frank, I.E. and Friedman,J.H. (1993). A Statistical View of Some Chemometrics Regression Tools. *Technometrics*, **35**, 109-148.

Franklin, C. H. (1989). Estimation across data sets: Two-stage auxiliary instrumental variables estimation. *Political Analysis*, **1**, 123.

Gajalakshmi, V. and Peto, R. (2004). Verbal autopsy of 80,000 adult deaths in Tamilnadu, South India. *BMC Public Health*, **4**.

Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (1995). *Bayesian Data Analysis*. London: Chapman & Hall.

Gelman, A. and Rubin, D.B. (1992). Inference from iterative simulations using multiple sequences with discussion. *Statistical Science*, **7**, 457-511.

Goldstein, H. (2002). Multilevel Statistical Models, 3rd ed. A Hodder Arnold Publication.

Hand, D. J. (2006). Classifier technology and the illusion of progress. *Statistical Science*, **21**, 114.

Heyde,C.C.(1994). A Quasi-likelihood approach to the REML estimating equation, *Statist and Probability Letters*, 21, 381-384.

Hopkins, D. and King, G. (2007). Extracting systematic social science meaning from text. Available at http://gking.harvard. edu/files/abs/words-abs.shtml.

Hoppa, R. D. and Vaupel, J. W., EDS. (2002). *Paleodemography*. Cambridge Univ. Press.

Jackman, S. (2000). Estimation and inference are missing data problems: unifying social science statistics via Bayesian simulations. *Political Analysis*, **8**, 307-332.

Jackman, S. (2001). Multidimensional Analysis of Roll Call Data via Bayesian Simulation: Identification, Estimation, Inference and Model Checking. *Political Analysis*, **9**, 227-241.

Jiang,J. (1996). REML estimations: Asymptotic Behavior and Related Topics, *Annals of Statistics*, 24, 255-286.

Jiang,J., and Rao,J.S. (2003), Consistent Procedures for Mixed Linear Model Selection, *Annals of Statistics*, 65, 23-42.

Kalter, H. (1992). The validation of interviews for estimating morbidity. *Health Policy and Planning*, **7**, 3039.

Krishna, A. (2008). Shrinkage-Based Variable Selection Methods for Linear Regression and Mixed-Effects Models. Dissertation, North Carolina State University.

Laird, N. M. and Ware, J. H. (1982). Random-Effects Models for Longitudinal Data. *Biometrics*, **38**, 963-974.

Levy, P. S. and Kass, E. H. (1970). A three population model for sequential screening for Bacteriuria. *American Journal of Epidemiology*, **91**, 148154.

Lopez, A., Ahmed, O., Guillot, M., Ferguson, B. D., Salomon, J. A., Murray, C. J. L. and Hill, K. H. (2000). *World Mortality in 2000: Life Tables for 191 Countries*. World Health Organization, Geneva.

Longregan, J. (2000). *Legislative Institutions and Ideology in Chile's Democratic Transition*. New York: Cambridge University Press.

Lu, Y. and McFarland, T. (2007). *Reveal Congressional Party Effects via Hierarchical Ideal Point Estimation*. paper presented at the 2007 Summer Political Methodology Meeting, Penn State University.

Lu, Y. and Wang, X. (2008). Software Manual for *IPE: R package for item response models with correlated structures*.

Mallow, C. L. (1973). Some comments on C_p. *Technometric*, 15, 661-675.

Martin, A.D. and Quinn, K.M. (2002). Dynamic ideal point estimation via Markov Chain Monte Carlo for the U.S. Supreme Court. *Political Analysis*, **10**, 134-153.

Martin, A.D., Quinn, K.M. and Lee Epstein. (2005). The 'Rehnquist' Court(?). *Law and Courts*, **15**, 18-23.

Martin, A.D., Quinn, K.M. (2008). MCMCpack, Markov chain Monte Carlo (MCMC) Package. Version 0.9-4. Available at http://mcmcpack.wustl.edu.

McCarty, N., Keith, T. and Rosenthal, H. (2001). The Hunt for Party Discipline in Congress. *American Political Science Review*, **95**, 673-687.

Mathers, C. D., Mafat, D., Inoue, M., Rao, C. and Lopez, A. (2005). Counting the dead and what they died from: An assessment of the global status of cause of death data. *Bulletin of the World Health Organization*, **83**, 171177.

Maude, G. H. and Ross, D. A. (1997). The effect of different sensitivity, specificity and cause-specific mortality fractions on the estimation of differences in cause-specific mortality rates in children from studies using verbal autopsies. *International Journal of Epidemiology*, **26**, 10971106.

Morris, S. S., Black, R. E. and Tomaskovic, L. (2003). Predicting the distribution of under-five deaths by cause in countries without adequate vital registration systems. *International Journal of Epidemiology*, **32**, 10411051.

Nishii, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *Annal of Statistics*, **12**, 758-765.

Pacque-Margolis, S., Pacque, M., Dukuly, Z., Boateng, J. and Taylor, H. R. (1990). Application of the verbal autopsy during a clinical trial. *Social Science Medicine*, **31**, 585591.

Poole, K.T. and Rosenthal, H. (1997). *Congress: A Political-Economic History of Roll Call Voting*. New York: Oxford University Press.

Poole, K.T. (2005). *Spatial models of parliamentary voting*. New York: Cambridge University Press.

Pu, W. and Niu X. (2006). Selecting mixed-effects models based on a generalized information criterion, *Journal of Multivariate Analysis*, **97**, 733-758.

Quigley, M. A., Chandramohan, D., Setel, P., Binka, F. and Rodrigues, L. C. (2000). Validity of data-derived algorithms for ascertaining causes of adult death in two African sites using verbal autopsy. *Tropical Medicine and International Health*, **5**, 3339.

Rao, C. R. and Wu Y. (1989). A strongly consistent procedure for model selection in a regression problem, *Biometrika*, **76**, 369-374.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment* tests. Copenhagen: Denmarks Paedagogiske Institute.

Rivers, D. (2003). Identification of Multidimensional Spatial Voting Models. *Unpublished Manuscript*.

Searle,S.R., Casella,G. and McCulloch,C.E. (1992). *Variance Components*. Wiley, New York.

Schwartz, G. (1978). Estimating the dimensions of a model, *Annals of Statistics*, 6, 461-464.

Schubert, G. (1974). *The judicial mind revisited*. Oxford University Press. London.

Segal, J.A. and Spaeth, H. J. (1997). *The Supreme Court and the Attitudinal Model*. Cambridge: Cambridge University Press.

Setel, P.W., Whiting, D. R., Hemed, Y., Chandramohan, D., Wolfson, L. J., Alberti, K. G. M. M. and Lopez, A. (2006). Validity of verbal autopsy procedures for determining causes of death in Tanzania. *Tropical Medicine and International Health*, **11**, 681696.

Setel, P. W., Sankoh, O., Velkoff, V. A., Mathers, C., Gonghuan, Y. et al. (2005). Sample registration of vital events with verbal autopsy: A renewed commitment to measuring and monitoring vital statistics. *Bulletin of the World Health Organization*, **83**, 611617.

Shibata, R. (1984). Approximate efficiency of a selection procedure for the number of regression variables. *Biometrika*, 71, 43-49.

Sibai, A. M., Fletcher, A., Hills, M. and Campbell, O. (2001). Non-communicable disease mortality rates using the verbal autopsy in a cohort of middle aged and older populations in Beirut during wartime, 198393. *Journal of Epidemiology and Community Health*, **55**, 271276.

Sireci, S., Wainer, H., and Thissen, D. (1991). On the reliability of testlet-based tests. *Journal of Educational measurement*, **28**, 237-247.

Sirovich, L. (2003). A pattern analysis of the second Rehnquist U.S. Supreme Court. *Proceedings of National Academy of Science*, **100**, 7432-7437.

Snyder, and Groseclose. (2000). Estimating Party Influence in Congressional Roll Call Voting *American Journal of Political Science*, **44**, 193-211.

Soleman, N., Chandramohan, D. and Shibuya, K. (2005). WHO Technical Consultation on Verbal Autopsy Tools. Geneva.

Soleman, N., Chandramohan, D. and Shilbuya, K. (2006). Verbal autopsy: Current practices and challenges. *Bulletin of the World Health Organization*, **84**, 239245.

Spaeth, H.J. (1979). *Supreme Court Policy Making*. San Francisco, Freeman.

Spaeth, H.J. (2001). *United States Supreme Court Judicial Database, 1953-2000*.

Sun, Y., W. Zhang and H., Tong (2007). Estimation of the covariance matrix of random effects in longitudinal studies, *The annals of Statistics*, **35**, 2795-2814.

Tanner, M.A. and Wong, W.W. (1987). The Calculation of Posterior Distributions by Data Augmentation. *Journal of the American Statistical Association*, **82**, 528-540.

Thisted, R. A. (1988). *Elements of Statistical Computing: Numerical Computation*. Chapman and Hall, New York.

Vander Vaart, A.M. (1998). *Asymptotic Statistics*, Cambridge University Press.

Wainer, H., Bradlow, E.T., and Wang, X. (2007). *Testlet Response Theory and Its Applications*. Cambridge University Press.

Wainer, H., and Kiely, G. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement*, **24**, 185-202.

Wainer, H., and Thissen, D. (1996). How is reliability related to the quality of test score? What is the effect of local dependence on reliability? *Educational Measurement: Issue and Practice*, **15(1)**, 22-29.

Yang, G., Rao, C., Ma, J., Wang, L., Wan, X., Dubrovsky, G. and Lopez, A. D. (2005). Validation of verbal autopsy procedures for adult deaths in China. *International Journal of Epidemiology*, **35**, 741748.

Yen, W. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, **30**, 187-213.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, B*, **68**, 49-67.

Zou, H. (2006). The Adaptive Lasso and its Oracle Properties. *Journal of the American Statistical Association*, **101**, 1418-1429.